세계인문학포럼
World Humanities Forum

제8회 세계인문학포럼

The 8th WORLD HUMANITIES FORUM 2025

프로그램북 B
: 분과회의 (AI)

PROGRAM BOOK B
: Parallel Sessions (AI)

2025. 11. 4(화) ~ 6(목)

4Tue - 6Thu Nov. 2025

안동국제컨벤션센터

Andong International Convention Center

Hall B, Room 201-203

주최 | 교육부 경상북도 GYEONGSANGBUK-DO 안동시 주관 | NRF 한국연구재단 추진기관 | 숭실대학교 Soongsil University 협력기관 | 국립경국대학교 GYEONGGUK NATIONAL UNIVERSITY 후원 | unesco UNDER THE PATRONAGE OF UNESCO 유네스코한국위원회

PROGRAM **CONTENTS**

PROGRAM **CONTENTS**

PROGRAM **SCHEDULE**

2025.11.04(화) DAY 1

09:00 - 09:30	등록		
09:30 - 12:00	분과회의 1		
	세션 1 201호 세션 5 207호	세션 2 202호 세션 6 208호	세션 3 203호 세션 7 209호 / 세션 4 206호 세션 8 210호
12:00 - 13:00	중식		
13:00 - 13:30	개회식		B1F B홀
13:30 - 15:00	기조강연 1	모하메드 알리 벤마크루프 모하메드 6세 폴리테크닉 대학교(UM6P) 교수	B1F B홀
15:00 - 15:30	휴식		
15:30 - 17:30	심포지엄 1	박태웅 의장 녹서포럼	201~203호
	심포지엄 2	아메드 나심 바르카위 학장 푸자이라 철학의 집 이희수 명예교수 한양대학교	B1F B홀
18:00 - 19:30	글로벌 네트워킹 디너 1		2F 로비

2025.11.05(수) DAY 2

09:00 - 09:30	등록		
09:30 - 12:00	분과회의 2		
	세션 1 201호 세션 5 207호	세션 2 202호 세션 6 208호	세션 3 203호 세션 7 209호 / 세션 4 206호 세션 8 210호
12:00 - 13:30	중식		
13:30 - 15:00	기조강연 2	권헌익 케임브리지대학교 트리니티 칼리지 석좌교수	B1F B홀
15:00 - 15:30	휴식		
15:30 - 17:30	심포지엄 3	피니스 찬탈랑시 인문사회 자문관 유네스코 방콕 사무소 석봉래 교수 앨버니아대학교	B1F B홀
	심포지엄 4	문유미 교수 스탠퍼드대학교 곽준혁 교수 중산대학교	201~203호
18:00 - 19:30	글로벌 네트워킹 디너 2		안동 시내

2025.11.06(목) DAY 3

09:00 - 09:30	등록		
09:30 - 10:10	안동학 특별강연	김언종 한국고전번역원 원장	B1F B홀
10:10 - 11:10	기조강연 3	염재호 태재대학교 총장	B1F B홀
11:10 - 12:00	폐회식		B1F B홀
12:00 - 13:00	중식		

2025.11.04 Tue. DAY 1

09:00 - 09:30	Registration		
09:30 - 12:00	Parallel Sessions 1		
	Session 1 Room 201 Session 5 Room 207	Session 2 Room 202 Session 6 Room 208	Session 3 Room 203 Session 7 Room 209 / Session 4 Room 206 Session 8 Room 210
12:00 - 13:00	Lunch		
13:00 - 13:30	Opening Ceremony		B1F Hall B
13:30 - 15:00	Keynote Speech 1	Mohammed Ali Benmakhlouf Professor, University of Mohammed VI Polytechnique(UM6P), Morocco	B1F Hall B
15:00 - 15:30	Break		
15:30 - 17:30	Symposium 1	Taewoong Park Chairman, Green Paper Forum	Room 201~203
	Symposium 2	Ahmed Nassim Barqawi Dean, Philosophy House, Fujairah, UAE Heesoo Lee Professor Emeritus, Hanyang University	B1F Hall B
18:00 - 19:30	Global Networking Dinner 1		2F Lobby

2025.11.05 Wed. DAY 2

09:00 - 09:30	Registration		
09:30 - 12:00	Parallel Sessions 2		
	Session 1 Room 201 Session 5 Room 207	Session 2 Room 202 Session 6 Room 208	Session 3 Room 203 Session 7 Room 209 / Session 4 Room 206 Session 8 Room 210
12:00 - 13:30	Lunch		
13:30 - 15:00	Keynote Speech 2	Heonik Kwon Chair Professor, Trinity College, University of Cambridge	B1F Hall B
15:00 - 15:30	Break		
15:30 - 17:30	Symposium 3	Phinith Chanthalangsy Regional Advisor for Social and Human Sciences, UNESCO Bangkok Office Bongrae Seok Professor, Alvernia University	B1F Hall B
	Symposium 4	Yumi Moon Professor, Stanford University Jun-Hyeok Kwak Professor, Sun Yat-sen University	Room 201~203
18:00 - 19:30	Global Networking Dinner 2		Andong Downtown

2025.11.06 Thu. DAY 3

09:00 - 09:30	Registration		
09:30 - 10:10	Special Lecture on Andong Studies	Eonjong Kim Director, Institute for the Translation of Korean Classics	B1F Hall B
10:10 - 11:10	Keynote Speech 3	Jaeho Yeom President, Taejae University	B1F Hall B
11:10 - 12:00	Closing Ceremony		B1F Hall B
12:00 - 13:00	Lunch		

I PROGRAM **DETAILS**

2025.11.04 Tue. DAY 1

09:00 - 09:30 등록

09:30 - 12:00 분과회의 1 (AI)

201호 1. AI 윤리

사회 **손화철 교수** 한동대학교

강연 **1-1. 정성훈 교수** 인천대학교
1-2. Barbara Janusz-Pohl(바르바라 야누시-폴) 교수 아담 미키에비츠대학교
1-3. 허유선 교수 경남대학교
1-4. Ho Manh Tung(호 만 퉁) 연구원 베트남 사회과학원 철학연구소

토론 **고인석 교수** 인하대학교
Michał Wawrzyńczak(마이클 바브진차크) 연구원 아담 미키에비츠대학교, 볼로냐 대학교

202호 2. AI와 인문학 연구

사회 **박건우 교수** 국립창원대학교

강연 **2-1. 김바로 교수** 한국학중앙연구원
2-2. Tony Veale(토니 비일) 교수 더블린대학교
2-3. 전준 교수 카이스트
2-4. Marcin Galiński(마르친 갈린스키) 교수 고르조프 비엘코폴스키 야곱 파라디스대학교

토론 **맹성현 교수** 태재대학교

203호 3. AI 시대의 교육(Roundtable)

사회 **이찬규 교수** 중앙대학교

강연 **3-1. 도성훈 교육감** 인천광역시교육청
3-2. 엄성우 교수 서울대학교
3-3. 김현철 교수 고려대학교
3-4. 김원중 교수 단국대학교

206호 4. AI와 번역

사회 **최호빈 교수** 국립경국대학교

강연 **4-1. 임영길 교수** 성균관대학교
4-2. Ke Hu(커 휴) 교수 맬버른대학교
4-3. 장요한 교수 계명대학교
4-4. Muhammet Emre Korkmaz(무함메트 에므레 코르크마즈) 교수 앙카라대학교

2025.11.04 Tue. DAY 1

09:30 - 12:00 분과회의 1 (공존)

207호 5. 뮤지움, 포용적 공간

사회 **박은경 교수** 동아대학교

강연 **5-1. 박성일 학예연구사** 서울대학교 규장각
5-2. Vicki Sungyeon Kwon(권성연) 학예연구사 로얄 온타리오 박물관
5-3. 김영희 학예연구사 국립중앙박물관
5-4. Masao Oi(오이 마사오) 교수 도시샤대학교

토론 **민병찬 교수** 충북대학교

208호 6. 한국의 문학과 세계

사회 **김태용 교수** 숭실대학교

강연 **6-1. 이승우 소설가** <한국문학>
6-2. Jean-Claude de Crescenzo(장-클로드 드 크레센조) 교수 엑스 마르세유대학교
6-3. 정용준 교수 서울예술대학교

209호 7. 디지털시대의 책과 도서관

사회 **문경훈 교수** 경상국립대학교

강연 **7-1. 김영욱 교수** 서울대학교
7-2. Guillaume Fau(기욤 포) 문헌실장 프랑스 국립도서관
7-3. Rafael Olea Franco(라파엘 올레아 프랑코) 교수 콜레히오 데 멕시코

토론 **우석균 교수** 서울대학교
차지연 교수 충남대학교

210호 8. 타자와의 만남

사회 **정세근 교수** 충북대학교

강연 **8-1. 임영진 교수** 전남대학교
8-2. Aditi Singh(아디티 싱) 교수 서울대학교
8-3. 서민규 교수 건양대학교
8-4. Suk Gabriel Choi(최석 가브리엘) 교수 타우슨대학교

토론 **조남호 교수** 국제뇌교육종합대학원대학교

12:00 - 13:00 중식

13:00 - 13:30 개회식

I PROGRAM **DETAILS**

2025.11.04 Tue. DAY 1

13:30 - 15:00	기조강연 1 (AI)
B홀	사회 **김선욱 명예교수** 숭실대학교 강연 **Mohammed Ali Benmakhlouf(모하메드 알리 벤마크루프)교수** 모하메드 6세 폴리테크닉대학교(UM6P) 인공지능과 인간 대화의 도전 토론 **김혜숙 회장** 국제철학연맹(FISP)
15:00 - 15:30	휴식
15:30 - 17:30	심포지엄 1: AI 거버넌스
201~203호	사회 **이찬규 교수** 중앙대학교 강연 **박태웅 의장** 녹서포럼 토론 **이중원 명예교수** 서울시립대학교 **이광호 교수** 한국교원대학교
15:30 - 17:30	심포지엄 2: 중동 지역의 공존
B홀	사회 **성일광 교수** 서강대학교 강연 **Ahmed Nassim Barqawi(아메드 나심 바르카위) 학장** 푸자이라 철학의 집 **이희수 명예교수** 한양대학교
17:30 - 18:00	이동
18:00 - 19:30	글로벌 네트워킹 디너 1

2025.11.05 Wed. DAY 2

09:00 - 09:30	등록
09:30 - 12:00	분과회의 2 (AI)
201호	**9. AI와 인간성** 사회 **이길산 교수** 경남대학교 강연 **9-1. 방종우 교수** 가톨릭대학교 **9-2. Balaganapathi Devarakonda(발라가나파티 데바라콘다) 교수** 델리대학교 **9-3. 보일(양성철) 소장** AI 부디즘 연구소 **9-4. Johnson Thomaskutty(존슨 토마스쿠티) 교수** 유니온 신학대학원 토론 **최성호 선임연구원** 서울대학교 아시아연구소
202호	**10. AI와 예술** 사회 **박평종 교수** 중앙대학교 강연 **10-1. 정서현 교수** 카이스트 **10-2. Alice Barale(앨리스 바라레) 교수** 밀라노대학교 **10-3. 박진완 교수** 중앙대학교 토론 **Wang Jiaqi(왕가기)** 고려대학교 박사과정
203호	**11. AI와 한국어 교육** 사회 **채영희 교수** 국립부경대학교 강연 **11-1. 이정희 교수** 경희대학교 **11-2. Satyanshu Srivastava(사티안슈 스리바스타바) 교수** 네루대학교 **11-3. 곽용진 대표이사** (주)이르테크 **11-4. Jieun Joe Kiaer(조지은) 교수** 옥스퍼드대학교 토론 **강현화 교수** 연세대학교
206호	**12. AI와 언어** 사회 **권만우 교수** 경성대학교 강연 **12-1. 이유미 교수** 중앙대학교 **12-2. Jean-Louis Vaxelaire(장-루이 박셀레르) 교수** 나뮈르대학교 **12-3. 박정원 교수** 한국외국어대학교 **12-4. Seok Kang(강석) 교수** 텍사스대학교 토론 **박진호 교수** 서울대학교
12:00 - 13:30	중식

I PROGRAM **DETAILS**

2025.11.05 Wed. DAY 2

09:30 - 12:00 분과회의 4 (공존)

207호

13. 전쟁과 공존

사회 **김형곤 교수** 건양대학교

강연 **13-1. 손경호 교수** 국방대학교
13-2. Miklós Zeidler(미클로스 자이들러) 교수 부다페스트대학교
13-3. 김지영 교수 숭실대학교
13-4. Linda Sunarti(린다 수나르티) 교수 인도네시아대학교

토론 **박제광 학예실장** 건국대학교 박물관

208호

14. 위기와 레질리언스

사회 **김기봉 명예교수** 경기대학교

강연 **14-1. 박혜정 전문연구원** 연세대학교
14-2. Frank Uekötter(프랑크 우에쾨터) 교수 보훔 루르대학교
14-3. 이세진 교수 호서대학교
14-4. Yoshiyuki Yama(야마 요시유키) 교수 간세이가쿠닌대학교

209호

15. 포스트콜로니얼 문화와 민족 문학

사회 **김태용 교수** 숭실대학교

강연 **15-1. 최성웅 박사** 읻다출판사
15-2. Melina Balcázar(멜리나 발카사르) 교수 콜레히오 데 멕시코
15-3. Kuwada Kohei(코헤이 쿠와다) 교수 도쿄대학교

210호

16. 뉴 노멀시대의 공존(Roundtable)

사회 **정원섭 교수** 경남대학교

강연 **16-1. 최훈 교수** 강원대학교
16-2. 박상혁 교수 동아대학교
16-3. 성신형 교수 숭실대학교
16-4. 정원섭 교수 경남대학교

12:00 - 13:30 중식

13:30 - 15:00 기조강연 2 (공존)

B홀

사회 **김선욱 명예교수** 숭실대학교

강연 **Heonik Kwon(권헌익) 석좌교수** 케임브리지대학교 트리니티 칼리지
인공지능과 사회적 영혼 사이

토론 **박명림 교수** 연세대학교

2025.11.05 Wed. DAY 2

15:00 - 15:30 휴식

15:30 - 17:30 심포지엄 3: AI 편향성

B홀

사회 **Jin Y. Park(박진영) 석좌교수** 아메리칸대학교

강연 **Phinith Chanthalangsy(피니스 찬탈랑시) 인문사회 자문관** 유네스코 방콕 사무소
Bongrae Seok(석봉래) 교수 앨버니아대학교

15:30 - 17:30 심포지엄 4: 동아시아의 공존

201~203호

사회 **양일모 교수** 서울대학교

강연 **Yumi Moon(문유미) 교수** 스탠퍼드대학교
Jun-Hyeok Kwak(곽준혁) 교수 중산대학교

17:30 - 18:00 이동

18:00 - 19:30 글로벌 네트워킹 디너 2

2025.11.06 Thu. DAY 3

09:00 - 09:30 등록

09:30 - 10:10 안동학 특별강연

B홀

강연 **김언종 원장** 한국고전번역원
안동의 간략한 역사와 인물들

10:10 - 11:10 기조강연 3

B홀

사회 **김선욱 명예교수** 숭실대학교

강연 **염재호 총장** 태재대학교
AI시대의 인간: 인간과 AI의 공진화(共進化: Co-Evolution)

11:10 - 12:00 폐회식

12:00 - 13:00 중식

I PROGRAM **DETAILS**

2025.11.04 Tue. DAY 1

09:00 - 09:30 Registration

09:30 - 12:00 Parallel Sessions 1 (AI)

Room 201

1. AI Ethics

Moderator **Whachul Son** Professor, Handong Global University

Presenters
1-1. Sunghoon Jung Professor, Incheon National University
1-2. Barbara Janusz-Pohl Professor, Adam Mickiewicz University
1-3. Eusun Heo Professor, Kyungnam University
1-4. Ho Manh Tung Researcher, Institute of Philosophy, Vietnam Academy of Social Sciences

Discussants
Insok Ko Professor, Inha University
Michał Wawrzyńczak Ph.D. Candidate, Adam Mickiewicz University, University of Bologna

Room 202

2. AI and Humanities Research

Moderator **Geonwoo Park** Professor, Changwon National University

Presenters
2-1. Baro Kim Professor, The academy of Korean Studies
2-2. Tony Veale Professor, University College Dublin
2-3. June Jeon Professor, KAIST
2-4. Marcin Galiński Professor, The Jacob of Paradies University in Gorzów Wielkopolski

Discussant **Sunghyon Myaeng** Professor, Taejae University

Room 203

3. Education in the Age of AI (Roundtable)

Moderator **Chankyu Lee** Professor, Chung-Ang University

Presenters
3-1. Seonghoon Do Governor, Incheon Metropolitan city office Professor of Education
3-2. Sungwoo Um Professor, Seoul National University
3-3. Hyeoncheol Kim Professor, Korea University
3-4. Wonjoong Kim Professor, Dankook University

Room 206

4. AI and Translation

Moderator **Hobin Choi** Professor, Gyeongguk National University

Presenters
4-1. Younggil Yim Professor, Sungkyunkwan University
4-2. Ke Hu Professor, The University of Melbourne
4-3. Yohan Jang Professor, Keimyung University
4-4. Muhammet Emre Korkmaz Professor, Ankara University

2025.11.04 Tue. DAY 1

09:30 - 12:00 Parallel Sessions 1 (Coexistence)

Room 207

5. Museums as Inclusive Spaces

Moderator **Eunkyung Park** Professor, Dong-A University

Presenters
5-1. Vicki Sungyeon Kwon Curator, Royal Ontario Museum (ROM)
5-2. Masao Oi Professor, Doshisha University
5-3. Seongil Park Curator, Kyujanggak Institute, Seoul National University
5-4. Younghee Kim Curator, National Museum of Korea

Discussant **Byoungchan Min** Professor, Chungbuk National University

Room 208

6. Korean Literature and the World

Moderator **Taeyong Kim** Professor, Soongsil University

Presenters
6-1. Seungwoo Lee Novelist, Magazine『Korean Literature』
6-2. Jean-Claude de Crescenzo Professor, Aix-Marseille Université
6-3. Yongjoon Jeong Professor, Seoul Institute of The Arts

Room 209

7. Books and Libraries in the Digital Age

Moderator **Kyunghoon Moon** Professor, Gyeongsang National University

Presenters
7-1. Guillaume Fau Book Curator, Bibliothèque Nationale de France
7-2. Rafael Olea Franco Professor, El Colegio de México
7-3. Younguk Kim Professor, Seoul National University

Discussants
Sukkyun Woo Professor, Seoul National University
Jiyeon Cha Professor, Chungnam National University

Room 210

8. Encountering the Other

Moderator **Segeun Jeong** Professor, Chungbuk National University

Presenters
8-1. Youngjin Yim Professor, Chonnam National University
8-2. Aditi Singh Professor, Seoul National University
8-3. Suk Gabriel Choi Professor, Towson University
8-4. Mingyu Seo Professor, Konyang University

Discussant **Namho Cho** Professor, University of Brain Education

12:00 - 13:00 Lunch

13:00 - 13:30 Opening Ceremony

I PROGRAM **DETAILS**

2025.11.04 Tue. DAY 1

13:30 - 15:00	Keynote Speech 1 (AI)	
B1F Hall B	Moderator	**Seon-Wook Kim** Professor Emeritus, Soongsil University
	Presenter	**Mohammed Ali Benmakhlouf** Professor, University of Mohammed VI Polytechnique, Morocco (UM6P) Artificial Intelligence and the Challenges of Human Conversation
	Discussant	**Heisook Kim** President, International Federation of Philosophical Societies
15:00 - 15:30	Break	
15:30 - 17:30	Symposium 1: AI Governance	
Room 201~203	Moderator	**Chankyu Lee** Professor, Chung-Ang University
	Presenter	**Taewoong Park** Chairman, Green Paper Forum
	Discussants	**Jungwon Lee** Professor Emeritus, University of Seoul **Kwangho Lee** Professor, Korea National University of Education
15:30 - 17:30	Symposium 2: The Coexistence in the Middle East	
B1F Hall B	Moderator	**Ilkwang Seong** Professor, Sogang University
	Presenters	**Ahmed Nassim Barqawi** Dean, Philosophy House, Fujairah, UAE **Heesoo Lee** Professor Emeritus, Hanyang University
17:30 - 18:00	Transfer	
18:00 - 19:30	Global Networking Dinner 1	

2025.11.05 Wed. DAY 2

09:00 - 09:30	Registration	
09:30 - 12:00	Parallel Sessions 2 (AI)	
Room 201	**9. AI and Humanity**	
	Moderator	**Gilsan Lee** Professor, Kyungnam University
	Presenters	**9-1. Jongwoo Bang** Professor, The Catholic University of Korea Songsin Campus **9-2. Balaganapathi Devarakonda** Professor, University of Delhi **9-3. Boil** Director, AI Buddhism Research Institute **9-4. Johnson Thomaskutty** Professor, The United Theological College (UTC), Bengaluru
	Discussant	**Seongho Choi** Senior Researcher, Seoul National University Asia Center
Room 202	**10. AI and the Arts**	
	Moderator	**Pyungjong Park** Professor, Chung-Ang University
	Presenters	**10-1. Seohyon Jung** Professor, KAIST **10-2. Alice Barale** Professor, University of Milan **10-3. Jinwan Park** Professor, Chung-Ang University
	Discussant	**Wang Jiaqi** Korea University Ph.D. Candidate
Room 203	**11. AI and Korean Language Education**	
	Moderator	**Younghee Chae** Professor, Pukyong National University
	Presenters	**11-1. Junghee Lee** Professor, Kyung Hee University **11-2. Satyanshu Srivastava** Professor, Jawaharlal Nehru University **11-3. Yongjin Kwak** CEO, IIR TECH **11-4. Jieun Joe Kiaer** Professor, University of Oxford
	Discussant	**Hyounhwa Kang** Professor, Yonsei University
Room 206	**12. AI and Language**	
	Moderator	**Mahnwoo Kwon** Professor, Kyungsung University
	Presenters	**12-1. Yumi Yi** Professor, Chung-Ang University **12-2. Jean-Louis Vaxelaire** Professor, Université de Namur **12-3. Jeongweon Park** Professor, Hankuk University of Foreign Studies **12-4. Seok Kang** Professor, The University of Texas at San Antonio
	Discussant	**Jinho Park** Professor, Seoul National University

I PROGRAM **DETAILS**

2025.11.05 Wed. DAY 2

09:30 - 12:00 **Parallel Sessions 2 (Coexistence)**

Room 207

13. War and Coexistence

Moderator **Hyunggon Kim** Professor, Konyang University

Presenters
13-1. Kyengho Son Professor, Korea National Defense University
13-2. Miklós Zeidler Professor, University of Budapest
13-3. Jiyoung Kim Professor, Soongsil University
13-4. Linda Sunarti Professor, University of Indonesia

Discussant **Jaegwang Park** Chief Curator, Konkuk University Museum

Room 208

14. Crises and Resilience

Moderator **Gibong Kim** Professor Emeritus, Kyonggi University

Presenters
14-1. Hyejeong Park Nesearch Fellow, Yonsei University
14-2. Frank Uekötter Professor, Ruhr University Bochum
14-3. Sejin Lee Professor, Hoseo University
14-4. Yoshiyuki Yama Professor, Kwansei Gakuin University

Room 209

15. Post-Colonial Culture and National Literature

Moderator **Taeyong Kim** Professor, Soongsil University

Presenters
15-1. Sungwoong Choi Ph.D., Itta Publishing (Director)
15-2. Melina Balcázar Professor, El Colegio de México
15-3. Kuwada Kohei Professor, The University of Tokyo

Room 210

16. Co-Existence in the Age of "The New Normal" (Roundtable)

Moderator **Wonsup Jung** Professor, Kyungnam University

16-1. Hoon Choi Professor, Kangwon National University
16-2. Sanghyuk Park Professor, Dong-A University
16-3. Shinhyung Seong Professor, Soongsil University
16-4. Wonsup Jung Professor, Kyungnam University

12:00 - 13:30 **Lunch**

13:30 - 15:00 **Keynote Speech 2 (Coexistence)**

B1F Hall B

Moderator **Seon-Wook Kim** Professor Emeritus, Soongsil University

Presenter **Heonik Kwon** Chair Professor, Trinity College, University of Cambridge
Between Artificial Intelligence and Social Soul

Discussant **Myunglim Park** Professor, Yonsei University

2025.11.05 Wed. DAY 2

15:00 - 15:30 **Break**

15:30 - 17:30 **Symposium 3: AI Bias**

B1F Hall B

Moderator **Jin Y. Park** Chair Professor, American University

Presenter **Phinith Chanthalangsy** Regional Advisor for Social and Human Sciences, UNESCO Bangkok Office
Bongrae Seok Professor, Alvernia University

15:30 - 17:30 **Symposium 4: Coexistence in East Asia**

Room 201~203

Moderator **Ilmo Yang** Professor, Seoul National University

Presenter **Yumi Moon** Professor, Stanford University
Jun-Hyeok Kwak Professor, Sun Yat-sen University

17:30 - 18:00 **Transfer**

18:00 - 19:30 **Global Networking Dinner 2**

2025.11.06 Thu. DAY 3

09:00 - 09:30 **Registration**

09:30 - 10:10 **Special Lecture on Andong Studies**

B1F Hall B

Presenter **Eonjong Kim** Director, Institute for the Translation of Korean Classics
A Brief History of Andong and Its Notable Figures

10:10 - 11:10 **Keynote Speech 3**

B1F Hall B

Moderator **Seon-Wook Kim** Professor Emeritus, Soongsil University

Presenter **Jaeho Yeom** President, Taejae University
Humanity in the Age of AI: Co-Evolution of Humans and AI

11:10 - 12:00 **Closing Ceremony**

12:00 - 13:00 **Lunch**

주제 소개
Forum Theme Introduction

“AI 대전환 시대의 인문학”
“공존을 위한 모색”

제8회 세계인문학포럼은 “AI 대전환 시대의 인문학”과 “공존을 위한 모색”을 공동주제로 선정하였다. 인문학은 변화하는 시대를 살피며 인간에 대한 근본적인 질문을 다시 던져 최선의 대답을 제시해야 한다.

AI는 이 시대에 패러다임 전환의 새로운 중심을 이루고 있다. AI 대전환은 우리의 사회와 개인의 삶을 근본적으로 바꾸어 놓고 있다. 다변적 세계의 본질, 인간적 삶의 가능 조건, 변화에 필요한 자질, 변화하는 시대의 윤리와 교육의 방향성, 인간 주체성과 정체성, 문화의 변화, 노동 시장의 변화에 따른 사회경제적 대변혁 등에 대해 인문학은 새로운 지평에서 응답해야 한다.

또한 지금은 자연과 문명의 공존, 문화 간의 공존, 세대 간의 공존, 기술과 인간의 공존 등 인류가 고민하는 많은 문제에 대한 담론이 더 구체화되어야 할 시점이다. 어제와 오늘을 돌아보는 데 그치지 않고, 인문학의 관점에서 내일을 향해 “우리가 무엇을 할 것인가”를 논의해야 한다.

THE 8th WORLD HUMANITIES FORUM
제8회 세계인문학포럼

주제 소개
Forum Theme Introduction

“The Humanities in the Age of AI Transformation”
“Exploring Paths to Coexistence”

The 8th World Humanities Forum has selected “The Humanities in the Age of AI Transformation” and “Exploring Paths to Coexistence” as its central themes. Together, they reflect the need for the humanities to revisit fundamental questions about human existence, values, and identity in the midst of rapid technological and societal change.

Artificial intelligence now stands at the heart of a new paradigm, reshaping the ways we think, work, and live. This transformation invites the humanities to engage deeply with questions such as the nature of a multi-layered world, the conditions for meaningful human life, the qualities required to adapt to change, and the evolving directions of ethics and education in the age of AI. It also calls for renewed reflection on human subjectivity and identity, cultural transformation, and the socioeconomic shifts brought about by changes in the labor market.

At the same time, the challenge of coexistence—between nature and civilization, among cultures, across generations, and between technology and humanity—demands a more tangible and forward-looking dialogue. The Forum aims to move beyond reflection on the past and present to ask, from a humanistic perspective, a vital question for our shared future: What can humanity do today to create a better tomorrow?

분과회의 세션 01
Parallel Session

AI 윤리
AI Ethics

AI로 생성된 이미지를 갖춘 버추얼 프로필 정체성의 문제

The Problems of Virtual Profile Identity with Generated Image by AI

정성훈
인천대학교 교수

Sunghoon Jung
Professor, Incheon National University

초록

기능적으로 분화된 사회에서는 마음과 사회의 체계적 간극으로 인해 개인의 정체성이 하나의 문제가 된다. 소셜미디어 기반 프로필 정체성은 언어와 자아의 우연성으로 인한 아이러니를 벗어날 수 없기 때문에 새로운 어휘를 필요로 하는 정체성 문제를 풀어가는 하나의 패러다임이며, 전통적 패러다임인 성실성과 진정성의 대안으로 부상하고 있다. 최근 프로필성은 인공지능 이미지 생성과 모션캡처 기술의 발전으로 '버추얼 프로필 정체성'이라는 새로운 국면으로 진입하고 있다. 이 정체성에서는 마음의 기반이 되는 첫 번째 몸(유기체)과 구별되는 마음과 사회 사이에 들어오는 두 번째 몸(몸의 이미지)이 등장한다. 버추얼 프로필 정체성의 기능은 통계학적 일반 동료의 관찰에 대한 이차 관찰을 가능하게 하는 것, 그리고 대안적 정체성 형성을 통해 잠재력을 실현하는 것이다. 하지만 그에 따라 새로운 구조적 커플링 문제가 생긴다.

Abstract

Self-identity becomes problematic due to systematic gap between mind and society in functionally differentiated society. Social media-based profile identity, unable to escape the irony stemming from the contingency of language and self, serves as a paradigm for evolving identity problem requiring new vocabulary. It is emerging as an alternative to the traditional identity paradigms of sincerity and authenticity. Recently, profilicity has entered a new phase called 'virtual profile identity' due to advancements in AI image generation and motion capture technology. A second body (body image) emerges, distinct from the first body (the organism) that forms the foundation of the mind, positioning itself between the mind and society. The functions of virtual profilicity enable second-order observation of statistical general peer's observation and realize potential through the formation of alternative identities. However, this gives rise to new structural coupling problems.

1. 도입

"너는 누구냐?" - "저는 김○○입니다"
"너의 이름이 무어냐고 묻지 않았다. 너는 누구냐?" - "저는 김□□와 박☆☆의 첫째 딸입니다."
"네 부모가 누구냐고 묻지 않았다. 너는 누구냐?" - "저는 △△중학교 학생입니다."
"네가 어느 학교에 다니는지 묻지 않았다. 네가 누구인지 물었다" - "저는 ……."

이 대화문은 미래엔 출판사에서 나온 중학교 도덕 교과서의 '자아 정체성' 단락의 처음에 나오는 6컷 만화의 대화 내용이다. 이 대화에서 묻는 자는 교통사고로 의식을 잃은 학생에게 들려온 목소리이므로 학생 자신의 또 다른 자아일 가능성이 높다. 즉 이 대화는 '나는 누구인가?(Who am I?)'라는 물음으로부터 이어지는 자문자답의 성격을 갖는다. 교과서에는 마지막 답변이 제시되어 있지 않다. 아마도 어떤 답변도 정답일 수 없기 때문일 것이다.

로티(Richard Rorty)의 표현을 빌자면, 언어의 우연성(contingency)과 자아의 우연성을 받아들이는 자, 낡은 어휘에서 벗어나 온전히 자기 자신의 것이 될 "마지막 어휘(last vocabulary)"를 만들어내려 하는 자는 아이러니스트(ironist)가 될 수밖에 없다.[1)]

'나는 **이다'라는 식으로 자아를 규정하는 어휘는 자아가 아무리 창조적인 메타포를 통해 다시 쓴다 하더라도,[2)] 언어는 사적 언어일 수 없기 때문에 그는 그 어휘가 다른 누군가의 모방이 아닐까 의심할 수밖에 없다. 그래서 우연적 자아 규정은 다시 부정될 수밖에 없다는 아이러니 혹은 패러독스에 빠지게 된다.

그런데 왜 교과서는 '중학생', '첫째 딸' 같은 역할 정체성을 참된 것으로 간주하지 않고 학생들로 하여금 계속 다시 자아 정체성을 묻도록 하는 것일까? 주어진 역할들을 성실하게 이행하는 것 또한 정체성 문제의 한 해결 방향일 것이며, 전통 사회에서 많은 개인들은 역할 정체성 이상의 것을 묻지 않는 경우가 많았을 것이다. 그런데 왜 현대 사회와 엮여서 살아가는 개인들은 패러독스에 빠지면서도 역할들과 동일시될 수 없는 진정한 정체성에 대한 물음을 던지는 것일까?

이 글은 사회와 개인의 관계에 대한 루만(Niklas Luhmann)의 설명을 통해 현대 사회에서 개인의 정체성 문제가 떠오르는 이유를 밝힌다. 그리고 아이러니를 벗어날 수 없기 때문에 계속 새로운 어휘와 이미지를 필요로 하는 정체성 문제를 풀어가는 하나의 패러다임으로서 소셜미디어 기반 프로필 정체성에 대해 살펴볼 것이다. 프로필 정체성이란 자아 정체성 문제를 "마음-사회 문제(mind-society problem)"[3)]로 규정하고, 성실성(sincerity)과 진정성(authenticity)으로부터 프로필성(profilicity)으로 정체성 패러다임이 바뀌고 있음을 고찰한 묄러(Hans-Georg Moeller)와 담브로시오(Paul J. D'Ambrosio)의 연구에 따른 것이다.

나는 여기서 더 나아가 프로필 정체성이 AI 이미지 생성으로 인해 새로운 국면에 진입하고 있으며 새로운 문제를 낳고 있다고 본다. 이 문제는 마음의 기반이 되는 첫 번째 몸(유기체)의 문제와는 구별되는 마음과 사

1) Rorty, Richard, Contingency, Irony, and Solidarity; 로티, 리처드, 『우연성, 아이러니, 연대』, 170쪽, 209쪽 등.
2) 로티는 이를 '메타포적 재기술(metaphoric redescription)'이라고 부른다. 위의 책 58.
3) Moeller, Hans-Georg & D'Ambrosio, Paul J., "Sincerity, authenticity and profilicity: Notes on the problem, a vocabulary and a history of identity", Philosophy and Social Criticism Vol. 45(5), 2019, 579.

회 사이에 들어오는 두 번째 몸(몸의 이미지)으로 인해 생기는 문제이다. '버추얼 프로필 정체성'의 문제라고도 표현할 수 있을 것이다.

'문제'라는 표현이 다소의 오해를 살 수 있기에 미리 언급하자면, 나는 프로필 정체성이나 그것의 버추얼화가 긍정적 기능과 부정적 후과를 동시에 갖는다고 본다. 프로필 정체성은 전통적인 정체성 패러다임을 통해서는 관찰하기 어려웠던 자신에 대한 이차 관찰을 가능하게 해주며, 대안적 정체성 형성을 통해 잠재력을 실현할 수 있게 해준다. 하지만 그에 따른 새로운 구조적 커플링 문제가 생기며, 이로 인해 신경증과 분열증이 커질 수 있다.

2. 현대 사회에서 개인의 정체성 문제

근대 철학이 인간과 관련해 마음(mind)과 몸(body)의 이원론을 기본으로 했다면, 루만의 체계이론은 인간의 몸과 마음을 구별할 뿐 아니라 사회에 참여하는 인간으로서의 인격(person)을 뚜렷이 나눈다. 그는 체계와 환경의 구별을 이용한 관찰을 통해 지칭할 수 있는 체계들을 기계들, 생명 체계들, 심리적 체계들, 사회적 체계들의 네 가지로 분류하며, 이 체계들이 각각 고유한 작동(operation) 방식을 갖고 있다고 본다.[4] 인간은 한편으로는 세포들의 자기생산(autopoiesis)[5] 네트워크를 기반으로 세포들 중 일부인 피부막을 통해 그것의 환경과 스스로 경계를 긋는 유기체(몸)이고, 다른 한편으로는 의식 작용들의 자기생산을 통해 그것의 환경과 경계를 유지하는 심리적 체계(마음)이다. 그에 반해 커뮤니케이션들의 자기생산을 통해 환경과의 경계를 긋는 사회적 체계들에서 인간의 몸과 마음은 그것들의 환경에 놓인다. 사회적 체계들에서 인간은 그 자체로는 작동적 단위가 아니며 커뮤니케이션이 귀속되는 단위인 인격으로서 관찰된다. 따라서 인격은 하나의 체계가 아니라 관찰에 의해 동일화되는 단위이며, 루만은 인격을 "커뮤니케이션 참여자들" [6], "개별 인간을 향한 기대들의 복합체를 사회적으로 동일화하는 지칭"[7] 등으로 규정한다.

자기생산적 체계들은 작동상의 폐쇄로 인해 그것들 고유의 작동들로는 결코 환경과 직접 접촉할 수 없기 때문에 환경에 있는 체계들과의 구조적 커플링(structural coupling)에 의존한다.[8] 지각할 수 없는 사회적 체계들은 공진화 매체인 언어를 통해 심리적 체계들과 구조적으로 엮여 있어서 마치 보고 있고 듣고 있는 듯한 문장들을 통해 환경에 '관한' 커뮤니케이션을 이어나갈 수 있다. 작동 자체로는 환경과 접촉할 수 없지만, 구조적 커플링을 통해 환경에 있는 체계들과 서로를 제약하면서, 즉 구조를 형성함으로써 환경에 관한 정보처리를 하는 것이다. 그래서 사회적 체계들은 심리적 체계들이 커뮤니케이션 참여에 동기유발될 수 있도록 기대구조를 형성해야 하며, 심리적 체계들 역시 커뮤니케이션 참여자가 되기 위해서는 사회적 기대구조에 맞춰 통지하거나 이해해야 한다. 이렇게 두 가지 종류의 체계들의 구조적 커플링이 별 잡음 없이 비교적 원활하게 이루어질 때 해당 참여자들은 인격들로 관찰된다.[9]

루만은 심리적 체계들이 구조적 커플링을 통해 인격이 되는 과정을 '사회화(socialization)'라고 부르며, 사회적 체계들 쪽에서 인간이 커뮤니케이션의 참여자 혹은 주소지로 간주되는 것을 '포함(inclusion)'이라고 부른다.[10]

이러한 사회화와 포함은 가족, 마을 등의 분절적 분화를 기초로 하여 중심/주변 분화, 계층적 분화 등의 분화 형식으로 추가된 전근대 사회들과 기능적 분화 형식이 주된 분화 형식이 되어 다른 분화 형식들을 부속화시킨 현대 사회에서 매우 다른 양상을 갖게 된다. 전통 사회들에서 사회화와 포함은 그가 속한 가족 혹은 지역적 공동체를 기초로 이루어졌다. 그리고 가족이나 공동체는 이미 사회의 부분체계들 중 하나에만 자리를 잡고 있었다. 예를 들어, 계층적으로 분화된 사회에서 한 사람이 귀족사회에 속하면서 동시에 농노사회에 속하는 것은 불가능했다. 간혹 귀족 임명 등의 신분 변화를 통해 포함된 부분체계들이 바뀔 수 있지만 대부분의 경우 그것 역시 개인의 성취는 아니었다. 이렇게 한 인격에게 할당된 부분체계가 뚜렷하게 규정되어 있기 때문에 개인은 자신에게 주어진 사회적 역할에 충실한 것말고는 다른 정체성을 가질 가능성이 극히 적었고 정체성 형성을 위한 특별한 노력도 필요하지 않았다.
그런데 기능적으로 분화된 현대 사회에서 개인의 사회화 과정에서는 교육의 기능이 커진다. 그리고 교육체계로의 포함은 가족과 무관한 개인으로서 이루어진다. 그리고 교육에서 시작된 경력(career)을 바탕으로 경제, 과학, 정치, 법 등의 기능체계들과 관련된 조직의 구성원이 되는 것 역시 개인 고유의 과제가 된다. 그런데 개인이 하나의 인격으로서 여러 가지 역할들을 맡을 수 있으려면 사회의 여러 부분체계들에 걸쳐서 살아가야 한다. 그래서 루만은 "각자가 오직 한 체계에만 속하는 식으로, 그러니까 법에만 참여하고 경제에는 참여하지 않는 식으로, 정치에만 참여하고 교육체계에는 참여하지 않는 식으로 인간들을 기능체계들에 할당할 수는 없다"고 말한다.[11]

현대적 개인은 하나의 부분체계로의 포함 혹은 그 부분체계의 조직의 구성원이 됨을 통해서는 그가 사회에서 차지하는 고유한 정체성을 확인할 수 없다. 그런 포함들을 통해 얻게 되는 것은 역할 정체성들일 뿐이고, 그런 역할들은 언제든지 타인들에 의해 대체될 수 있는 것이다. 그래서 개인의 심리적 체계는 여러 사회적 체계들과의 커플링 문제, 즉 여러 기대구조들에 맞추는 문제에 시달리게 된다. 그리고 그는 정해진 프로그램에 따른 역할 수행의 시간이 지나고 나면 '나는 누구인가?'라는 물음을 던지게 되곤 한다. 현대 사회에서는 심리적 체계가 사회의 환경에 놓여 있음이 뚜렷해지며, 여러 역할 정체성들과는 다른 개인적 정체성을 추구하려는 경향이 커진다.

그런데 도입부에서도 지적했듯이 '나는 누구인가?'라는 물음에 대한 답은 논리적으로 볼 때는 '나는 나다'라는 공허한 동어반복(tautology)일 수밖에 없고, 술어를 무엇으로 규정하는 순간 자아 정체성은 패러독스에 빠지게 된다. 예를 들어 '나는 학생이다'에 대한 긍정은 다른 학생들도 많다는 사실과 내가 영원히 학생일 수는 없다는 사실로 인해 부정될 수 있다. 역할이 아닌 다른 규정들, 예를 들어 '키가 크다', '배려심이 많다' 등에 대한 긍정 역시 마찬가지로 부정될 수 있다. 그래서 술어에 무엇을 집어넣어도 자아 정체성에 대한 규정은 긍정과 부정 사이를 진동할 수밖에 없는 패러독스가 된다.

이러한 자아 정체성의 패러독스를 은폐하는, 혹은 패러독스를 다른 방향으로 옮겨놓는 관계나 방법 역시 등

4) uhmann, N., Soziale Systeme – Grundriß einer allgemeinen Theorie, Frankfurt am Main: Suhrkamp, 1984, 16.
5) 루만은 마투라나와 바렐라가 생명체를 설명하기 위해 도입한 자기생산 개념을 심리적 체계들과 사회적 체계들로 확장해 사용한다. 그는 자기생산 체계를 "자신을 이루는 요소들을 바로 그 요소들 자체의 네트워크를 통해 산출하는 체계"라고 규정한다.
6) Luhmann, N., Die Gesellschaft der Gesellschaft, Frankfurt am Main: Suhrkamp, 1997, 106; 장춘익 역, 『사회의 사회』, 새물결, 2014, 133.
7) Luhmann, N., Soziale Systeme, 286.
8) Luhmann, N., Die Gesellschaft der Gesellschaft, 92-100; 『사회의 사회』 117-127.
9) Luhmann, N., Die Gesellschaft der Gesellschaft, 106; 『사회의 사회』 133.
10) Luhmann, N., "Wie is Bewußtsein an Kommunikation beteiligt?", Soziologische Aufklärung 6, 2.Auflage, 2005, 51. 영역본은 "How Can the Mind Participate in Communication", Theories of Distinction: Redescribing the Descriptions of Modernity, Stanford: Stanford University Press, 2002.
11) Luhmann, N., Die Gesellschaft der Gesellschaft, 744; 『사회의 사회』 854.

장한다. 연애로 대표되는 친밀관계, 참된 정체성 형성 모델에 대한 독서 등이 그러하다. 그런데 나의 고유한 세계를 확인해주는 관계, 전인격에 관해 커뮤니케이션하는 관계로 간주되는 친밀관계 역시 하나의 사회적 체계일 뿐이며 여기서도 정직성에 대한 의심이나 커뮤니케이션불가능성의 경험은 이미 18세기의 문학에서 지적되었다.[12] 또한 정체성 모델에 대한 학습은 결국 모방이기에 고유하지 않다는 패러독스에 빠진다. 그런데 많은 개인들이 나는 누구인지를 물으며, 패러독스에 빠지면서도 계속 새로운 규정을 시도한다. 로티의 표현으로는 아이러니스트가 되는 것이고, 루만의 표현으로는 패러독스의 전개(Entfaltung der Paradoxie), 즉 패러독스를 펼쳐나간다.[13]

3. 프로필 정체성의 기능과 문제

정체성 패러다임에 대한 묄러와 담브로시오의 연구는 루만의 관점, 즉 현대적 정체성이 매우 불안정하다는 것으로부터 출발한다. 그들은 마음과 사회 사이에는 "체계적인 간극(systemic gap)"이 있으며, 기능적 분화의 조건 아래서 "사회적 정체성"은 유동적이고 다양하며 서로 합치되지 않는다고 본다.[14] 묄러와 담브로시오는 '자아(self)', '페르소나(persona)', '정체성(identity)'에 대한 그들 나름의 개념 정의로부터 논의를 시작한다.

그들은 '자아'를 개인의 심리적 체계를 뜻하는 용어로 사용한다. 따라서 자아는 생각과 느낌 속에서 경험하는 나 자신으로 규정된다. '페르소나'는 루만의 인격(person) 개념과 비슷한 것으로, 즉 커뮤니케이션의 개인적 주소지이자 여러 역할들의 담당자로 규정된다. 그리고 묄러와 담브로시오는 오늘날 페르소나가 유튜브, 페이스북, 트위터 등 소셜미디어 계정(account)을 갖는다는 점을 덧붙인다. 마지막으로 그들은 '정체성'이라는 용어를 "자아가 스스로와 동일화할 수 있고 하나의 페르소나가 그것과 동일화될 수 있는 물리적, 정신적(mental), 사회적 구성물"이라고 규정한다. 그들은 정체성이 개인의 자기 개념화이자 하나의 '투사(projection)'이며, 이 투사는 사회 안에서의 '인정(recognition)'을 찾기 위한 것이라고 말한다.[15] 이렇듯 투사가 사회적 인정을 바라기 때문에 정체성은 어려운 문제가 된다.

묄러와 담브로시오는 근대 유럽에서의 정체성 문제를 탐구한 트릴링(Lionel Trilling)의 1972년 저서 Sincerity and Authenticity를 참조해, 전근대 사회 혹은 근대 초기의 직업 중심 사회에서 지배적이었던 정체성 패러다임을 '성실성'과 '헌신(devotion)'으로 규정한다. 성실성과 헌신은 오늘날에도 학교생활, 직장생활 등에서 강조되는 정체성이다. 그리고 그들은 사회적 이동성이 증가한 현대 사회에서 "고유한 본래적 자아"를 추구하는 정체성을 '진정성'이라고 부른다. 그리고 로티를 참조하여 진정한 자아의 우연성과 아이러니를 드러낸다.

묄러와 담브로시오는 트릴링이 언급하지 않았던 새로운 정체성 패러다임을 제시한다. 그들은 후기 현대 혹은 소셜미디어의 시대에 "이차 관찰을 위해 제시되는 우리 자신의 이미지"이자 "자기 자신에 대한 간접적 관찰"로 형성되는 정체성을 '프로필성(profilicity)' 혹은 '프로필 정체성'이라고 부른다.[16] 루만과 사이버네틱스 이론가들이 자주 쓰는 개념인 '이차 관찰(second-order observation)'은 일차 관찰자가 사용하는 구별을 관찰하는 것, 즉 일차 관찰자의 맹점을 관찰하는 것이다. 성실성 패러다임과 진정성 패러다임에서 개인은 대체로 자기 자신에 대한 일차 관찰에 머무른다. '나는 학생이므로 열심히 공부해야 한다'는 관찰에는 학생/비학생 구별이 적절한가에 대한 의심이 허용되지 않는다. 그리고 진정성은 사회적 역할들과 타인의 시선을 배제하고 주관적인 내면 관찰을 추구한다는 점에서 역시 일차 관찰의 성격을 갖는다. 그에 반해 프로필성은 타인들의 관찰을 고려해 자아를 제시하고 그에 대한 타인들의 반응을 참조해 다시 자아를 관찰한다는 점에서 일차 관찰자의 맹점을 관찰하는 이차 관찰의 성격을 갖는다.

고프먼(Erving Goffman)의 일상 생활에서의 '자기-제시(self-presentation)'[17] 에 대한 연구에서 볼 수 있듯이, 이미 대면 상호작용에서도 개인들은 사회적 관계와 상황에 맞추어 자신을 드러내고 평판을 고려해 이미지를 관리했다. 소셜미디어의 도입 이전에도 자기 자신에 대한 이차 관찰 혹은 프로필성은 어느 정도 있었던 것이다. 프로필이라는 용어 자체가 예전부터 '한 측면에서 본 이미지' 혹은 '개인의 공식 경력을 서술한 문서' 등의 뜻으로 쓰였다. 하지만 묄러와 담브로시오는 소셜미디어가 주된 매체가 되면서 과거의 공식 프로필과 달리 스스로가 타인의 관찰을 고려해 편집할 수 있는 프로필성과 이를 통해 이루어지는 이차 관찰이 예전보다 강화되었다고 본다. 특히 자기 제시가 직접 아는 사람들이 아니라 통계 데이터로 등장하는 '일반 동료(general peer)'를 겨냥하게 되었다는 점에서 큰 차이가 있다.

'일반 동료'란 루소의 '일반 의지' 개념, 그리고 오늘날 학술지의 브라인드 '동료 심사(peer review)'가 같는 익명성을 참조해 묄러와 담브로시오가 만들어낸 표현이다. 구체적으로 누구인지는 알 수 없지만 알고리듬을 통해 통계적으로 떠오르는 타인들이 일반 동료이다.[18] 그래서 소셜미디어의 프로필을 큐레이팅하는 개인은 수많은 일반 동료의 관찰을 겨냥하여 프로필 이미지를 수정하고 피드에 노출될 게시물을 기획한다. 그리고 '좋아요', '구독', '공유', '슈퍼챗' 등 일반 동료의 관찰에 대한 통계 데이터를 관찰해 이것들을 수정하고 편집한다. 즉 이차 관찰을 통해 프로필 정체성을 재구성한다.

대면 상호작용과 언어적 상호작용을 중시하는 사람들은 프로필성을 참된 개성의 상실이나 소외로 간주하기도 한다. 예를 들어, 디지털 매체 시대의 정체성을 '스크린 자아'로 규정하면서 묄러와 담브로시오의 프로필 이론을 검토한 신정아와 최용호는 프로필 기반 정체성의 테크놀로지를 비판적으로 분석한 후 다음과 같은 결론을 내린다. "우리 시대 스크린 자아는 자신의 개성을 드러내는 것이 아니라 자신의 프로필을 드러낸다. 개성을 지배하는 정체성 테크놀로지가 언어라면 프로필을 지배하는 정체성 테크놀로지는 숫자다. 스크린 자아의 개별화 원리는 숫자의 지배를 받는다. 숫자에 의한 정체성 거버넌스는 우리 시대가 새로운 소외의 국면으로 접어들고 있음을 시사한다."[19]

그런데 이들이 설정하는 대립구도, 즉 '개성'과 '언어'를 하나로 묶어서 '프로필'과 '숫자'에 대립시키는 구도는 과연 적절한가? 소셜미디어 플랫폼이 통계 숫자를 제시하며 이차 관찰이 이에 대해 민감하게 반응하는 것은 사실이지만, 소셜미디어에서 이루어지는 커뮤니케이션에서도 여전히 언어의 비중은 높다는 점에서 언어 대 숫자의 대립구도는 그리 적절해 보이지 않는다. 이미지와 글로 이루어진 페이스북 게시물에 '좋아요'

12) Luhmann, N., Liebe als Passion, Frankfurt/M.: Suhrkamp, 1982; 정성훈 외 역, 『열정으로서의 사랑』, 새물결, 2009.

13) Luhmann, N., "Individuum, Individualität, Individualismus", 228.

14) Moeller, Hans-Georg & D'Ambrosio, Paul J., "Sincerity, authenticity and profilicity: Notes on the problem, a vocabulary and a history of identity", 579.

15) Moeller, Hans-Georg & D'Ambrosio, Paul J., 581.

16) Moeller, Hans-Georg & D'Ambrosio, Paul J., You and Your Profile – Identity After Authenticity, Columbia University Press, 2021, 10-16.

17) 한국어 번역본의 영향으로 '자아 연출'로 번역되는 경향이 있는데, 나는 presentation을 연극의 performance와 구별하기 위해 '제시'로 번역한다. 고프먼, 어빙, 『자아 연출의 사회학』, 현암사, 2016 참조.

18) Moeller & D'Ambrosio, You and Your Profile, 47-50.

19) 신정아·최용호, 「스크린 자아: 디지털 미디어 시대의 정체성」, 『프랑스어권 문화예술연구』 제90집, 2024, 169.

를 누르는 일반 동료 중에는 이미지만 보고 대충 누르는 경우도 있겠지만 제법 많은 수의 사람들은 여전히 그 글을 읽고 이해한 후에 누른다. 유튜브 실시간 방송을 보다가 구독을 누르는 일반 동료 중 많은 수는 댓글 창을 매개로 이루어지는 유튜버와의 채팅을 좋아한다. 그래서 프로필 정체성의 테크놀로지를 비언어적이라고 간주하는 것은 적절치 않다. 다만 숫자의 영향력이 커졌다는 점에는 주목할 필요가 있다. 그리고 숫자에 대한 집착으로 생기는 부작용에 대해서는 분명히 주목할 필요가 있다.

뮐러와 담브로시오는 프로필 큐레이팅에 집착하다가 생기는 '프로필 노이로제' 문제를 제기한다. 그들은 통계에 대한 관찰에 지쳐 일정 기간 방송을 중단한 유명 요가 유튜버의 사례를 든다.[20] 물론 그 유튜버는 휴식 후에 다시 방송을 재개했다. 한국의 많은 유명 유튜버들도 활동 중단과 복귀를 반복하곤 한다. 프로필 정체성에 집착할수록 피로와 노이로제는 피하기 어렵다. 그래서 뮐러와 담브로시오는 프로필 정체성 패러다임 속에서 살아가는 개인들에게는 간혹 소셜미디어로부터 퇴각해 제정신(sanity)을 차리는 시간이 필요하다고 말한다.[21]

프로필과 개성을 대립시키는 것은 프로필이 아닌 참된 개성이 별도로 있다는 관점에서 나오는 대립구도이다. 그런데 뮐러와 담브로시오가 성실성, 진정성, 프로필성 중 하나의 정체성만 택하고 나머지는 버려야 말하는 것은 아니다. 게다가 그들이 진정성을 아이러니로 간주한다는 점을 고려하면, 프로필 정체성은 대면 공간에서의 전통적 역할들과 더불어 자아 정체성의 패러독스를 펼치는 기능을 한다고 보는 것이 적절하다. 개인은 자아 정체성을 고민하거나 추구할 때 때로는 자신의 역할 수행을 중심으로 한 관찰을 통해, 때로는 프로필 통계에 대한 관찰을 통해, 때로는 비사회적 시공간으로의 퇴각을 통해 정체성 문제를 풀어나갈 수 있다. 그들이 쓴 책의 결론부에서 뮐러와 담브로시오는 "우리는 아침에 진정성 있게 깨어날 수 있고, 낮 동안에 성실하게 우리의 일을 할 수 있고, 밤에는 우리의 공개 프로필을 큐레이팅할 수 있다"[22] 고 말한다. 그런데 깨어날 때의 진정성이란 사실 그들이 말한 '혼돈'[23] 의 시간, 즉 정체성을 추구하지 않는 시간이라고 보아야 할 것이다. 깨어나서 모바일폰을 보거나 타인의 얼굴을 보는 순간부터 우리는 혼돈에서 벗어나 다양한 자기 제시를 해야 하며, 이런 자기 제시를 위해 특정한 정체성 패러다임 혹은 정체성 모드를 선택해야 한다. 그래서 우리의 삶은 끊임없이 정체성의 패러독스를 풀어가는 과정으로 보아야 하며, 그 과정에서 프로필 정체성 또한 이런 패러독스 전개에 중요한 기능을 한다.

4. 버추얼 프로필 정체성의 부상

앞서 보았듯이 뮐러와 담브로시오는 정체성을 "물리적, 정신적, 사회적 구성물"로 규정함으로써 체계적 간극이 있는 몸, 마음, 페르소나를 하나로 투사하고 인정받고자 하는 것으로 설정했다. 그런데 다양한 메타버스의 등장, 그리고 소셜미디어에서 신체를 가리거나 변형할 수 있는 동영상 커뮤니케이션 기법이 발전함으로써, 최근에는 이런 동일성(identity) 혹은 구조적 커플링이 다시 나누어져 재편성되는 경향이 나타나고 있다. 이미 메타버스의 아바타, 디지털 게임의 부캐 등 개인이 디지털 가상 공간에서 자신의 두 번째 정체성을 형성하는 일은 흔히 일어나고 있다. 그래서 아바타들 사이에 일어나는 버추얼 성폭력이나 부캐를 이용해 타인의 게임 아이템을 약탈하는 일[24] 등은 이미 십여 년 전부터 사회적 이슈로 떠올랐다.

그런데 최근에는 예외적이거나 일탈적으로만 볼 수 없는 경향들이 나타나고 있다. 개인이 자신의 본체와 뚜렷이 구별되는 버추얼 프로필 정체성을 통해 잠재력을 실현하고 많은 일반 동료가 팬덤을 형성하는 일이 대중문화 내부에서 독자적인 영역으로 자리잡아 가고 있다. 본체가 이미 드러나 있는 연예인들의 부캐 활동이 가장 널리 알려진 사례이지만, 아예 전적으로 디지털 버추얼 이미지만으로 등장해 활동하는 유튜버나 스트리머들도 점차 큰 인기를 얻고 있다.

AI를 통한 이미지 생성과 모션캡처 기술의 발전은 이 새로운 영역의 발전과 대중화에 큰 영향을 미치고 있다. 자신의 얼굴을 드러내지 유튜버들은 과거에 카메라 방향 조절, 가면 쓰기 등의 기법을 썼다. 기획사 등의 도움으로 세련된 버추얼휴먼 이미지를 쓰는 경우에는 많은 제작비용이 들었을 뿐 아니라 '불편한 골짜기(uncanny valley)' 문제로 인해 시청자의 외면을 받기도 했다. 젊은 세대에게 친숙한 만화 이미지를 사용하는 것이 점차 확산되었는데, 이 또한 그림을 잘 그려야 했을 뿐 아니라 얼굴 표정과 신체 움직임을 동영상으로 표현하기 위해서는 고가의 모션캡처 장비와 스튜디오가 필요했다. 특히 여러 개의 센서가 달린 수트를 입고 촬영하면 그것을 자동으로 애니메이션으로 바꾸어주는 모션캡처 기술은 몇 년 전까지만 하더라도 큰 제작비가 투여되는 영화나 게임에서 사용되었고 연예기획사에 소속된 버추얼아이돌만 쓸 수 있었다. 그런데 생성형 AI가 사용자가 요구하는 만화 이미지를 생성해주고, 아이폰만 있으면 스튜디오와 특수 장비 없이도 모션캡처를 가능하게 하는 AI 서비스[25]가 시작되면서, 이제 미술적 재능이 없거나 소속사가 없는 개인 유튜버들도 적은 비용으로 버추얼 프로필 정체성을 구축할 수 있게 되었다.

만화 이미지의 캐릭터로 모션캡처를 통해 방송을 하는 유튜버들은 이 기법의 선구자인 일본의 키즈나 아이(Kizuna Ai)가 2016년에 처음 등장할 때 쓴 자신에 대한 규정을 따라서 보통 '버추얼유튜버(virtual youtuber, v-tuber)'라고 불리며, '버튜버'라고 약칭되기도 한다. 한국에서 가장 활발하게 활동하고 있는 버추얼유튜버는 아이돌 가수 활동을 병행하는 '이세계아이돌', '플레이브' 등이 있다. 이들이 연예기획사에 소속되어 고가 장비의 도움으로 활동하는 반면에, 최근에는 저가의 모션캡처 서비스 덕택에 소속사 없는 개인 버튜버도 엄청나게 늘어나고 있다. 그들 중 다수는 노래와 춤을 중심을 활동하지만, 간혹 그림 그리기, 책 읽기 등 다양한 영역에서 자신의 본체를 가리고 방송을 진행한다.

과거에도 연예기획사의 투자로 이른바 '얼굴 없는 가수'나 애니메이션 이미지를 이용한 '버추얼휴먼 가수'가 시도되었다. 하지만 최근의 버튜버가 그들의 구독자 혹은 팬덤과 맺는 관계에서 나타나는 새로운 특징은 디지털 버추얼 신체와의 구조적 커플링을 통해 버추얼 프로필 정체성 자체를 거의 독자적으로 인격화하고 있다는 것이다. 그런 특징 몇 가지를 살펴보도록 하자.

얼굴 없는 가수나 버추얼휴먼 가수의 팬덤은 가수의 본래 얼굴을 무척 궁금해하면서 밝히고자 노력했다. 하지만 최근 버추얼유튜버의 팬덤은 대부분 디지털 매체를 통해 드러난 페르소나에 집중할 뿐 그것의 본체에

20) Moeller & D'Ambrosio, You and Your Profile, 180.

21) Moeller & D'Ambrosio, 229.

22) Moeller & D'Ambrosio, 226.

23) 그들은 『장자』 제7편 7장에 나오는 '혼돈의 죽음' 이야기를 정체성 레짐의 등장으로 해석한다. 남해의 신과 북해의 신이 중앙의 신인 혼돈의 얼굴에 구멍을 뚫어주자, 즉 정체성을 부여하자 혼돈은 죽어버렸다는 이야기이다. 참고로 뮐러는 루만 체계이론 입문서를 쓴 사람이기도 하지만 원래 노자와 장자를 연구한 동양철학자이다.

24) '부캐'는 2000년대 초반 온라인 게임 플레이어들이 두 개 이상의 계정을 사용하면서 생겨난 신조어이다. 처음에는 일종의 반칙에 해당하는 것이었기 때문에 상당히 부정적인 의미로 쓰였다. 한 명의 플레이어가 본캐릭터와 부캐릭터의 두 개의 캐릭터를 만들어서 부캐로 획득한 아이템을 본캐가 가져가거나 제재를 당한 본캐 대신 부캐로 온갖 변칙 플레이를 하는 데 쓰곤 했다. 그런데 2018년 '쇼미더머니'에 한 래퍼가 복면을 쓴 채 '마마손'이라는 부캐로 등장해 본캐와는 다른 스타일의 무대를 선보인 이후 부캐는 잠재력 발휘라는 긍정적 의미 또한 갖게 되었다.

25) 아이폰과 연동해 사용하는 모션캡처 프로그램인 Move AI의 경우 유료 personal plan 요금이 현재 1년에 126달러이다. Chat GPT 유료 버전보다 저렴한 것이다.

대해서는 무관심하다. 버튜버 개인의 과거나 본체에 대해 이른바 "빨간약", "전생 공개" 등의 제목으로 간혹 게시물이 뜨기도 하지만, 팬덤은 그런 폭로가 확산되지 않도록 막는 데 적극적으로 나선다. "캐릭터와 인간을 구별하지 않아요. 제가 좋아하는 것은 그냥 그 하나의 자체예요"[26] 라는 플레이브 팬덤의 인터뷰 내용은 모니터 화면의 배후에 있는 인간에 대해 더 이상 관심이 없다는 것을 뜻하며, 버튜버 팬덤은 대부분 이런 태도를 보인다.

그리고 아이돌 가수 활동을 하는 버추얼유튜버는 다른 아이돌들에 비해 활동 공간에 제약이 많다. 그래서 실시간 커뮤니케이션이 가능한 플랫폼 공간에 더욱 집중하며, 거의 매일같이 실시간 방송을 진행하기도 한다. 이런 커뮤니케이션 체계에서는 실제 신체가 아닌 디지털 이미지가 사태적 동일성으로 사용되는 기대구조의 형성이 이루어지고 있다. 따라서 버추얼 프로필 정체성이 인격화되고 있다고 볼 수 있다. 그리고 버추얼유튜버의 인격은 적어도 디지털 동영상 커뮤니케이션의 기대구조에서는 아예 본래의 신체와 구조적으로 커플링된 것으로 등장하지 않는다는 점에서 인격의 버추얼화라고 부를 수도 있을 것이다. 물론 인쇄 시대에도 저자들의 신체는 드러나지 않았다. 그런데 오직 글로만 등장하는 인격과 달리 버추얼화된 인격은 본체와는 다른 이미지의 신체 운동을 동반하면서 등장하고 원활한 구어 상호작용을 한다는 점에서 큰 차이를 갖는다. 버튜버의 팬덤은 연예기획사가 버튜버의 페르소나가 가진 고유한 개체성을 훼손하는 일에 대해 저항하기도 한다. 버튜버의 원조인 키즈나 아이의 기획사는 성우 한 명으로는 감당하기 어려울 정도로 아시아권 전역에 걸쳐 활동 수요가 늘어나자 2019년에 일본어 성우 두 명과 중국어 성우 한 명을 추가 투입해 네 명의 성우가 하나의 캐릭터를 연기하게 했다. 그런데 키즈나 아이의 고유성을 지키려는 팬들이 불만을 터뜨리며 불매운동을 벌이고 일부 성우가 말실수를 하는 등 여러 문제가 생기자 결국 다시 원래의 성우 한 명이 담당하게 되었다.[27] 이 때문인지 한국의 인기 버튜버인 이세계아이돌의 경우 데뷔할 때부터 본체의 가수와 같은 것으로 짐작되는 나이, 고향, 예전 직업 등을 밝혔다. 그래서 이세계아이돌 멤버들의 본명이나 본체를 알아내는 것은 그리 어렵지 않지만 놀랍게도 팬덤은 그에 대해 철저하게 무관심하다.

버추얼유튜버는 버추얼 프로필 정체성이 인격화되는 추세를 보여주는 하나의 사례일 뿐이다. 최근 소셜미디어에서는 자신의 프로필을 생성형 AI를 이용해 만든 지브리 스타일 만화 이미지로 대체하는 것이 크게 유행했던 것처럼, 많은 사람들이 자신의 얼굴과 신체를 변형하여 커뮤니케이션 참여자가 되려고 한다. 그 이유가 무엇이었던지[28] 정체성을 이루는 성분들 중 어떤 것, 특히 얼굴 이미지를 변형하여 소셜미디어의 일반 동료들에게 보이고자 하는 경향은 늘어나고 있으며, 이를 통해 신체를 드러내는 상호작용에서는 불가능했던 잠재력을 실현하기도 한다.

5. 새로운 구조적 커플링과 인격화의 문제

버추얼 프로필 정체성이 부상하고 있다는 것은 이러한 정체성 모드가 자아 정체성 문제를 풀어가는 데 있어서 이차 관찰, 잠재력 실현 등 여러 가지 기능(function)을 하기 때문일 것이다. '기능적 분석'을 연구 방법으로 사용하는 루만에 따르면, 하나의 문제에 대해서는 하나의 해법만 있는 것이 아니며, "문제와 문제풀이의 관계(Die Relation von Problem und Problemlösung)"[29] 를 분석함으로써 여러 가지 기능적 등가물(funktionales Äquivalent)을 찾아낼 수 있다. 버추얼 프로필 정체성 역시 본래의 신체 이미지로는 해결하기 어려웠던 문제를 풀어나가는 하나의 기능적 등가물이다. 그런데 기능한다는 것은 곧 그 기능으로 인한 문제를 낳는다. 경제, 정치, 법 등의 기능체계들이 환경파괴, 양극화, 공동체 파괴 등의 문제들을 낳듯이, 대안적 정체성 패러다임 또한 기능함으로써 문제를 일으킨다. 여기서는 버추얼 프로필 정체성으로 인해 생기는 문제 두 가지를 다루도록 하겠다.

첫 번째 문제는 새로운 구조적 커플링이 낳는 문제이다. 앞서 보았듯이 프로필 정체성에 관한 연구에서 묄러와 담브로시오의는 정체성을 '마음-사회 문제'라는 관점에서 접근하였다. 그들은 루만처럼 몸은 이미 마음과 구조적으로 커플링되어 있는 것으로 간주했다. 마음의 인프라로서 몸을 전제했을 뿐, 사회의 커뮤니케이션들에서 등장하는 몸의 이미지 효과에 대해서는 독자적으로 다루지 않았다. 특히 몸짓과 얼굴 표정 등의 제스처 커뮤니케이션, 즉 비언어적 커뮤니케이션의 양상 변화에 대해서는 주목하지 않았다. 물론 그들의 책에는 일반 동료의 관찰을 겨냥하여 편집되는 소셜미디어의 프로필 사진들에 대한 수많은 언급들이 있지만, 그런 정체성 이미지가 아예 버추얼 이미지로 대체되는 경우는 고려하고 있지 않다.

프로필 정체성에서 몸은 마음의 인프라인 유기체(본체)로만 전제되거나 비록 보정과 편집을 거치긴 하지만 유기체의 복제본으로만 간주된다. 그에 반해 버추얼 프로필 정체성에서는 AI 이미지 생성이라는 기계적 프로세스에 의존하는 두 번째 몸이 등장한다. 이 정체성을 유지하려는 개인은 유기체를 움직이는 것과 화면에 나타난 몸의 이미지에 동시에 집중해야 한다. 컴퓨터와 네트워크의 성능 한계로 인해 본체의 움직임이 애니메이션에 늦게 반영되는 이른바 동기화(synchronization) 실패, 촬영 현장의 여건으로 인해 아예 몸의 일부가 화면에서 사라져버리는 사고 등이 일어나기 때문이다. 잘 훈련된 가수는 자신의 첫 번째 몸을 제어하여 노래하고 춤추는 데서는 크게 어려워하지 않는다. 커뮤니케이션을 겨냥하여 몸-마음의 구조적 커플링을 원활하게 하는 것이다. 그런데 기계적 시스템에 의존해야 하는 두 번째 몸, 즉 보이는 몸을 제어하는 일은 쉬운 문제가 아니다. 두 번째 몸과의 구조적 커플링까지 원활하게 이루어져야 그는 버추얼 프로필 정체성을 형성할 수 있다.

이러한 새로운 구조적 커플링의 문제는 프로필 노이로제를 증폭시킬 뿐 아니라 분열증을 심화시킬 수 있다. 노이로제와 마찬가지로 분열증은 하나의 페르소나로 여러 가지 역할들을 돌아가면서 수행해야 하는 현대적 자아 누구나 어느 정도 겪는 것이다. 그런데 페르소나마저 거의 이원화되는 버추얼유튜버에게 이 분열증은 더 커질 수밖에 없다. 더구나 애니메이션 캐릭터 이미지가 본체 이미지와의 괴리가 클 경우 분열증은 심화될 수밖에 없다. 버튜버들 중에는 자신을 용궁의 황녀, 조선시대 선비 등으로 설정하거나 나이를 1700세로 설정하는 등 상당히 무리한 신비주의 캐릭터 설정을 하는 경우도 있는데, 이런 경우 과연 장기간 활동이 가능할지 의문스러우며 실제로 힘겨움을 호소하며 활동을 중단한 사례들도 있다.

두 번째 문제는 페르소나의 이원화로 인해 일어날 수 있는 문제, 즉 버추얼 프로필 정체성의 법적 지위 문제이다. 이미 키즈나 아이의 사례는 버추얼유튜버의 캐릭터를 그저 상표권으로 간주해도 되는지의 문제를 던졌다. 그 권리를 연예기획사에 귀속시키느냐 성우에게 귀속시키느냐라는 상표권 분쟁의 문제는 그들 사이에 맺어진 계약의 공정성 문제이다. 따라서 아티스트 본인의 권리와만 관련된 문제이므로 새로운 문제는 아

26) 오윤지, 「버추얼 아이돌 팬덤의 향유 문화 연구」, 『한국콘텐츠학회논문지』 Vol.24, No.2, 2024, 189.

27) 글로벌이코노믹 이원용 기자 입력(2022-02-27 13:11), "버추얼 유튜버 '대모' 키즈나 아이, 6년만에 잠정 은퇴". https://www.g-enews.com/article/ICT/2022/02/202202262243408182c5fa75ef86_1

28) 사생활 보호, 직장생활 병행, 딥페이크 피해 예방, 나이와 외모로 인한 편견 극복 등 여러 가지 이유를 짐작해 볼 수 있겠지만, 실증적 연구의 과제이므로 여기서는 논하지 않겠다.

29) Luhmann, N., Soziale Systeme, 84.

니다. 새로운 문제는 그 캐릭터에 대해 독자적으로 형성된 커뮤니케이션의 기대 복합체를 그저 '사물'로 간주할 수 있는가이다. 계약상의 권리 문제와는 별도로 팬덤이라는 하나의 사회적 체계는 캐릭터와 결합된 목소리의 본인의 교체를 허락하지 않을 수도 있고, 어쩌면 그 본인에게 문제가 생겼을 경우 그 캐릭터를 가장 잘 이어갈 수 있는 다른 방법을 고안해낼지도 모른다. 상상적인 예측을 하나 해보자면, 목소리와 채팅 내용을 학습한 AI를 이용해 구현하고자 노력할 수도 있다.

이런 예측은 더 이상 캐릭터로 부르기 어려운 수준으로 준인격화된 페르소나를 인격과 사물의 이분법에 따라 그저 사물로 간주하는 것이 적절한가라는 의문을 불러 일으킨다. 이미 메타버스의 아바타의 법적 지위와 관련해서는 몇몇 연구성과가 나왔고, 대체로 법학자자들은 아바타의 독자적인 법적 인격성 인정에 대해서는 부정적인 견해를 보이고 있다.[30] 그리고 가상 인간(버추얼휴먼)의 법적 지위를 다각도로 연구한 김현귀는 가상 인간의 캐릭터를 저작권, 퍼블리시티권 등 재산권적으로 보호하는 데는 한계가 있다고 보면서 앞으로 인격권 인정에 관해 논의해볼 수 있다고 말한다. 하지만 그가 설정하는 '가상 인간'은 배후에 본체에 해당하는 개인이 없는 경우이다.[31]

나 역시 캐릭터의 배후에 권리를 가진 본인, 즉 인간-인격이 분명히 있는 프로필 정체성에 대해서는 독자적인 법인격을 논하는 것은 부적절하다고 본다. 하지만 본체의 인간-인격과 구별될 수 있는 기대 복합체의 정체성을 그저 인격이 아니라는 이유로 재산권의 대상으로 삼는 것이 적절한지에 대해서는 의문을 던질 수밖에 없다. 이것은 지금처럼 법이 계속 인격과 사물의 이원론을 고수해야 하는지에 대한 의문이다. Pietrzykowski가 지적하듯이 이제 우리에게는 "사물성(thinghood)과 인격성(personhood) 사이의 실재(reality)"[32] 를 다룰 세분화된 새로운 언어와 새로운 법적 개념이 필요한 것으로 보인다. 인공지능의 법적 지위 논쟁과도 연결될 수 있는 이 논의에 관해서는 향후의 과제로 남겨두고 여기서는 문제만 제기하고자 한다.

30) 이영록, 「메타버스 아바타의 법적 지위 -인격성 인정 여부를 중심으로-」, 『서울법학』 제31권 제2호, 2023, 1-38.

31) 김현귀, 「가상인간의 인격에 대한 법적 고찰」, 『미디어와 인격권』 제8권 제3호, 2022, 1-42.

32) Pietrzykowski, Tomasz, "The Idea of Non-personal Subjects of Law", Legal Personhood: Animals, Artificial Intelligence and the Unborn, Springer, 2017.

참고문헌

고프먼, 어빙, 『자아 연출의 사회학』, 현암사, 2016.
김현귀, 「가상인간의 인격에 대한 법적 고찰」, 『미디어와 인격권』 제8권 제3호, 2022.
신정아·최용호, 「스크린 자아: 디지털 미디어 시대의 정체성」, 『프랑스어권 문화예술연구』 제90집, 2024.
오윤지, 「버추얼 아이돌 팬덤의 향유 문화 연구」, 『한국콘텐츠학회논문지』 Vol.24, No.2, 2024.
이영록, 「메타버스 아바타의 법적 지위 -인격성 인정 여부를 중심으로-」, 『서울법학』 제31권 제2호, 2023.
정창우 외, 『2022 개정 중학교 도덕 ①』, 미래엔, 2025.
Luhmann, Niklas, Liebe als Passion, Frankfurt/M.: Suhrkamp, 1982; 정성훈 외 역, 『열정으로서의 사랑』, 새물결, 2009.
Luhmann, Niklas, Soziale Systeme – Grundriß einer allgemeinen Theorie, Frankfurt am Main: Suhrkamp, 1984.
Luhmann, Niklas, Die Gesellschaft der Gesellschaft, Frankfurt am Main: Suhrkamp, 1997; 장춘익 역, 『사회의 사회』, 새물결, 2014.
Moeller, Hans-Georg & D'Ambrosio, Paul J., "Sincerity, authenticity and profilicity: Notes on the problem, a vocabulary and a history of identity", Philosophy and Social Criticism Vol. 45(5), 2019.
Moeller, Hans-Georg & D'Ambrosio, Paul J., You and Your Profile – Identity After Authenticity, Columbia University Press, 2021.
Pietrzykowski, Tomasz, "The Idea of Non-personal Subjects of Law", Legal Personhood: Animals, Artificial Intelligence and the Unborn, Springer, 2017.
Rorty, Richard, Contingency, Irony, and Solidarity, Cambridge University Press, 1989; 김동식·이유선 역, 『우연성, 아이러니, 연대』, 사월의책, 2020.

형사사법에서의 윤리적 쟁점과 AI 도구의 적용 – 윤리 규범의 이중성 및 판결 휴리스틱에 대한 논의

Ethical Issues and AI Tools Application in Criminal Justice – Discussion of Duality of Ethical Norms and Adjudication Heuristics

바르바라 야누시-폴
아담 미키에비츠대학교 교수

Barbara Janusz-Pohl
Professor, Adam Mickiewicz University

Abstract

This work examines the ethical implications of AI applications to automate judicial decision-making. We will highlight the importance of ethics, including the axiological standards that a legal framework for AI should uphold. We will also point out that the legal framework must be characterised by a certain 'openness'. Nonetheless, the purely ethical standard is essential. We will focus on the types of AI systems that automate court decisions, and then demonstrate that not only the choice of AI instrument but also specific behavioral patterns based on heuristics can influence its true significance for the judge's cognitive processes. The backdrop to this discussion reflects a tension between the pursuit of efficiency through the implementation of AI technologies and the fundamental principles of human rights (i.e., the right to court).
The paper consists of three sections. In the first part, we will outline a general discussion on the ethical risks of using artificial intelligence in the criminal justice system. Next, we will focus on the impact of ethical standards on the shape of AI regulations, citing several examples, such as the concept of Trustworthy AI. This concept assumes that AI must be constantly monitored in legal, ethical, and praxeological contexts. In the third part of this paper, we will focus on the practical aspects of using AI tools to support the issuance of court decisions. We will draw attention to various models, as well as heuristic risks associated with disrupting the judge's cognitive process.

Introduction

The integration of Artificial Intelligence (AI) into criminal justice systems presents a complex landscape of opportunities. AI systems are most commonly used in advanced case law search engines, online dispute resolution, legal document drafting assistance, predictive analytics, automated compliance verification, and legal aid chatbots. When it comes to criminal justice, across the globe, jurisdictions are actively integrating AI tools at various stages of criminal proceedings. These advanced technologies include predictive policing algorithms that identify and forecast high-crime areas, as well as sophisticated risk-assessment tools that inform crucial bail and sentencing decisions. As these innovations become more prevalent, AI is progressively emerging in criminal justice, promising efficiencies but also raising critical ethical considerations (Gans-Combe, 2020; Said G, Azamat K, Ravshan S, Bokhadir A., 2023).[1)]

This article examines the ethical implications of applying AI to automate judicial decision-making (Zhang, Chen, &Song, 2024). We will highlight the importance of ethics, including the axiological standards that a legal framework for AI should uphold. We will also point out that the legal framework must be characterised by a certain 'openness'. Nonetheless, the importance of purely ethical standards, i.e., normative structures that are not introduced into binding legal systems as legal norms, is essential. We will focus on the types of AI systems that automate court decisions. We will demonstrate that not only the choice of AI instrument, but also heuristic aspects, can influence its true significance for the cognitive processes of the judge.

The backdrop for this discussion reflects a tension between the pursuit of efficiency through the implementation of artificial intelligence technologies and the fundamental principles of human rights that underpin axiological issues such as fairness, transparency, and the preservation of due process rights, particularly in the context of the right to defence.

Significantly, in the criminal procedural law system, efficiency aimed at reducing costs and the duration of proceedings does not exhaust the entire concept of effectiveness (efficacy). The latter is understood in relation to achieving the objectives of criminal proceedings, which also includes achieving axiologically based objectives (Janusz-Pohl, 2025B). One shall recall that among the goals of the proceedings is to promote material justice and crime prevention, to recognise the rights of the injured party, and finally to ensure efficiency in the economic dimension. This means that efficiency, streamlining, and acceleration are not values in themselves, but must be constantly balanced against other objectives of the process that

1) See more: The Council of Europe Study DGI(2017)12, Algorithms and Human Rights– Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, prepared in March 2018 by the Committee of Experts on Internet Intermediaries (MSI-NET), March 2018; https://rm.coe.int/algorithms-and-human-rights-en rev/16807956b5 [accessed on 5 October 2025].

are guarantee-oriented and axiological. The axiological context is therefore considered to take precedence over efficiency. Unfortunately, the values underlying the principles of criminal procedure do not always form a coherent catalogue, and some of them may be somewhat contradictory (Janusz-Pohl, 2022).

When analysing the implementation of the standard of protection of fundamental rights in criminal proceedings, it is necessary to note that the central figure in the proceedings (in this regard) is the defendant. He or she is protected by the right to defence, which broadly encompasses, inter alia, the ability to confront evidence, the presumption of innocence, and the principle of equality before the law. The literature highlights that the "black box" nature of many AI algorithms has raised significant concerns about how these algorithms may infringe upon defendants' constitutional rights to due process and a fair trial, as they often lack transparency, making it difficult for defendants to contest their validity. It is essential to strike a balance between the effectiveness of the criminal justice system and the rights of the accused. Currently, many defendants do not have full access to the methodologies and algorithms underlying the AI-driven evidence or risk assessments used against them.

Research has shown that AI systems can produce inconsistent and biased outcomes across different demographic groups, which exacerbates inequalities within the justice system and raises questions about the fairness that should be inherent in the legal process (Said, Azamat Ravshan & Bokhadir, 2023) This lack of transparency not only undermines the presumption of innocence but also raises significant concerns regarding equal protection under the law. When considering the use of AI in criminal proceedings, it is also essential to address procedural opportunism—situations in which criminal prosecution may be limited or abandoned. In these cases, the role of algorithms in decision-making could increase significantly. Precisely because of the aforementioned central position of the accused, it turns out that the issue of using AI (sometimes associated with a certain potential threat to the exercise of the right of defence) may depend on whether a given instrument is used in favour of or to the disadvantage of the accused. Legal systems recognise many institutions that allow for adjudication with a significant restriction of the scope of the right to defence (so called abbreviated procedures), provided that the adjudication is in favour of the defendant or, at least, that the given formula results in some benefits for the defendant compared to a "full" procedure, and the defendant is guaranteed a quick and informal review of the final decision (Janusz-Pohl, 2023; Janusz-Pohl, Wawrzyńczak, 2023).

When examining AI systems that play a pivotal role in decision-making, it becomes crucial to systematically classify and evaluate how these automated technologies may inadvertently embed biases and obscure the reasoning behind decisions that significantly affect individuals' lives. Achieving a harmonious balance between the remarkable capabilities of technological advancements and the fundamental principles of ethics is vital to ensuring fair and equitable treatment for all. The scholarly literature reveals a spectrum of pressing ethical challenges, including algorithmic bias, the opacity of decision-making processes, and the potential erosion of human judgment in critical legal determinations (Said, Azamat, Ravshan, Bokhadir, 2023; Zafar, 2024).

Furthermore, there is a concerning drift away from an anthropocentric perspective when technology takes precedence. Our examination will focus on the ethical implications of employing artificial intelligence in adjudicative functions, where the consequences extend far beyond merely arriving at a decision based on the merits of a case. We assert that the adjudication process should encompass the 'entire cognitive journey of the court', weaving together the threads of analysis, reasoning, and human insight in a comprehensive and thoughtful way.

2. Humans and technology
- normative context and dualistic nature of ethical norms

When analysing the ethical issues of the use of artificial intelligence in criminal justice, we must start with the fundamental problem of the relationship between humans and technology. In this context, it is worth noting that scholarly literature articulates the assumption of an anthropocentric focus that should guide the relationship between humans and AI (Floridi, 2018). It is recognised that the development of artificial intelligence that supports humans is unavoidable and generally beneficial when skillfully exploited, while limiting the risks associated with its use (Floridi, Cowls, Beltrametti, Chatila, Chazerand, Dignum, Luetge, Madelin, Pagallo, Rossi, Schafer, Valcke, Vayena, 2018). Limiting risks requires specific actions, the core of which is the creation of an adequate legal framework. Since artificial intelligence is a relatively new creation, it requires recognition and the establishment of a conventional world in which the rules for its use are defined. Although these issues are relatively new, ethical standards are crucial for shaping the legal framework. Thus, at the outset of our work, let us explore how ethics is reflected in normative, legally binding documents on AI.

Let it be immediately added that the development of such a legal framework is not an effortless undertaking. At the global level, these are primarily soft law instruments (not purely sources of binding legal norms, but rather 'formal embodiments' of ethical standards). An example of such a document is **the UNESCO Recommendation on the Ethics of Artificial Intelligence, adopted on November 24, 2021.**[2)] This document marks the first comprehensive global framework aimed at ensuring the ethical governance of AI technologies. As mentioned before, the recommendation outlines ten fundamental principles and identifies eleven key

2) SHS/BIO/PI/2021/1, Available at Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library [accessed on 1 Oct. 2025].

areas of action for effectively regulating AI applications. Central to these principles is the assertion that the design and implementation of AI systems must reflect four essential values: respect for human rights, protection of individual freedoms, enhancement of human dignity, and advancement of social justice and open justice.

When referring to the ethical aspect in normative terms, we must undoubtedly also turn to European regulations (namely, those of the Council of Europe and the European Union). However, let us pay attention to the methodological issue, because, on the one hand, the ethical aspect gains a 'thetic' justification when regulated by binding law. Still, on the other hand, some ethical elements remain in the sphere of (only) ethical norms, at most in the form of recommendations and guidelines (soft law), and therefore without a strong so-called 'thetical justification' (Ziembiński, 2021). In the European context, where the use of artificial intelligence has been regulated (see below AIA) **axiological aspects have been incorporated into the structure of legal norms**. This does not, however, mean that legal regulations have negated the importance of purely ethical approaches (those that fall outside the legal framework) to the use of AI. On the contrary, they are relevant given the dualistic nature of ethical norms (autonomous ethical norms and ethical norms conjoined with legal norms).

In this context, attention should be paid to the first document, namely the European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment, adopted on 3– 4 December 2018 in Strasbourg by the European Commission for the Efficiency of Justice (CEPEJ) of the Council of Europe during its 31st plenary meeting.[3] It is the first European text to set out ethical principles for the use of artificial intelligence (AI) in judicial systems, providing a framework to guide policymakers and legislators. Based on the Charter, the application of AI in the justice system can improve efficiency and quality. It must be implemented in a responsible manner which complies with the fundamental rights guaranteed, in particular, in the European Convention on Human Rights (ECHR) and the Council of Europe Convention on the Protection of Personal Data, consequently set of principles contains: the principle of respect for fundamental rights, the principle of non-discrimination, the principle of quality and security, the principle of transparency, impartiality and fairness, and the principle of control. It should be added that when discussing the fundamental rights approach based on the Charter on the Use of Artificial Intelligence in Judicial Systems, the most important and initial issue is to ensure the right to the court in automated judicial proceedings and ensuring the integrity of the entire procedure (level of formal justice) - art. 6 European Convention on Human Rights. When it comes to the right to a fair trial, the nuances of e-courts should therefore be considered. There is no doubt that the possible introduction of "adjudicating artificial intelligence" could have a positive impact on the speed of proceedings. Still, the question is about the cost and how to ensure participants a fair trial, i.e., what is referred to as procedural justice. As is well known, many researchers argue that, from the perspective of the person concerned by a given decision, the manner in which it is made is sometimes more important than its content. This includes both procedural guarantees and the communication process that takes place in court (Janusz-Pohl, 2025B). Another principle referred to in the Charter is non-discrimination, which encompasses three fundamental countermeasures: the creation of multidisciplinary research teams tasked with designing systems without discriminatory trends; the identification of discrimination during the system's operation; and the implementation of remedial measures. Charter indicates guidance that the processing of judicial decisions and data should be (praxeological implications).

From the perspective of our considerations on the ethical aspects of AI, which encompass only ethical standards that have not been transformed into binding legal norms, documents such as **the Ethics Guidelines for Trustworthy AI, a document issued publicly on April 8 2019 by** the High-Level Expert Group on AI (AI HLEG an independent expert group that the European Commission set up in June 2018)[4] are crucial. They allow us to define those patterns of behaviour that are key to the current shape of the body of ethical norms that have not (yet) been awarded by thetic justification, or for which such justification[5] is unnecessary. We can therefore see that the challenges associated with transferring specific competences to artificial intelligence, which until now have been rooted in human activity, require in-depth reflection on what the concept of trustworthy AI entails. **The idea encompasses three key areas: legality (pertaining to legal norms), axiology (concerning ethical norms), and praxeology (a set of instrumental propositions that serve as tools to explore this ideal model).** It is said expressis verbis in the Ethics Guidelines for Trustworthy AI, that trustworthy AI encompasses three components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. As indicated in the document *'Each component in itself is necessary but not sufficient for the achievement of Trustworthy AI. Ideally, all three components work in harmony and overlap in their operation. If, in practice, tensions arise between these components, society should endeavour to align them'.*

Certain ethical assumptions are explicitly stated in the document. It is recognised that the reference concept of trustworthy AI is based on the principles of respect for human autonomy,

3) CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment - European Commission for the Efficiency of Justice (CEPEJ)

4) Ethics Guidelines for AI [accessed on 1 Oct. 2025].

5) Provisory character of some of them is underlined in the statement of the members of the AI HLEG that "they do not necessarily agree with every single statement in the document" [first page of the document].

prevention of harm, fairness, and explicability. When searching for axiological assumptions, it is directly indicated that the ethical basis for trustworthy AI is corpus human rights within a framework of democracy and the rule of law, by reference to dignity, freedoms, equality and solidarity, citizens' rights and justice. The common foundation for these remains human dignity, or more specifically, a "human-centric approach," in which the human being enjoys a unique and inalienable moral status of primacy in the civil, political, economic, and social fields. The allocation of functions between humans and AI systems should follow human-centric design principles, leaving meaningful opportunities for human choice. This means securing human oversight. It should support humans in the workplace and aim to create meaningful work.

In the analysed document, we can also find the seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability. Attention is also drawn to the praxeological aspect, as one of the recommendations is to implement a Trustworthy AI assessment list when developing, deploying or using AI systems. Furthermore, one assumption is that such an assessment list will never be exhaustive. It is therefore emphasised that the concept of trustworthy AI constitutes a continuum. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders.[6)]

Moreover, in the document, it is said that *'Achieving Trustworthy AI requires not only compliance with the law, which is but one of its three components. Laws are not always up to speed with technological developments, can at times be out of step with ethical norms, or may simply not be well-suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms'.* This statement highlights a certain dualism in the ethical aspect, sometimes referring to the axiological content of legal norms and at other times to purely ethical models of conduct. Let us add that in the presented concept of trustworthy AI, the third element, which complements the other two, is the pragmatic aspect. Therefore, it is recognised that *'Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should operate in a safe, secure, and reliable manner, and safeguards should be implemented to prevent any unintended adverse impacts. It is therefore important to ensure that AI systems are robust. This is needed both from a technical perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase), and from a social perspective (in due consideration of the context and environment in which the system operates)'.*

The presented assumption of the dualistic influence of axiology, firstly in the form of applicable regulations, and secondly in the form of ethical standards, and additionally also the praxeological aspect – practical standards of 'good conduct', is reflected in the regulations adopted in the EU, namely Regulation (EU) 2024/1689 of the European Parliament and of the Council of June 13 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 Artificial Intelligence Act (AIA).[7)] It sets out the allowable uses of AI by categorising them according to the level of risk they pose to users and the population. As a side note, it should be noted that the AIA favours a risk-based approach over a rights-based approach and is intended to establish "common normative standards". The AI Act classifies AI according to its risk: 1) Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI). 2) high-risk AI systems, which are regulated. 3) A smaller section handles limited risk AI systems, subject to lighter transparency obligations; 4) Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI-enabled video games and spam filters. It is worth noting that, from the perspective analysed later in this paper — i.e., issues related to the ethical aspects of automating judicial decisions — we operate under the section on high-risk AI systems, which are supposed to be regulated. Since the indicated regulation is complex and very detailed, and the scope of this study is limited, we must refrain from a comprehensive discussion. At the same time, it should be noted that from the perspective of our considerations, **the AIA is an example of the recognition of specific ethical standards as legal standards;** the AIA imposes an obligation on the harmonisation of national legislation on member states; however, the regulations contained in the AIA are also part of the legal order of member states. Specific ethical principles are highlighted not only in individual articles but also in the preamble itself.

3. Selected problems – automation of adjudication and heuristics

So far, we have analysed the ethical aspect of the use of artificial intelligence **in an abstract way, i.e. in relation to the influence of ethics on the formation of regulatory (legal) models.** Still, it seems that the autonomy and the broad scope of purely ethical standards connected with AI — standards that remain complementary to legal norms (the regulatory sphere) — reflect the dynamics of AI development.

Now, we are moving to entirely new territory in terms of AI ethical context, and thus its applicability, in particular focusing on **the importance of ethics for AI application to the automation of decision-making in court proceedings,** namely to the AI methods able to

6) See: the Ethics Guidelines for Trustworthy AI.

7) PE/24/2024/REV/1; OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/oj

conduct legal reasoning required to make a judgment in judicial proceedings, so knowledge-based AI systems, machine learning systems or combination of these methods (Fernández-Martínez & Fernández, 2020).

As we have indicated earlier, the exercise of judicial functions in practical scenarios involves a range of activities, from analysing documents to supporting decision-making by judges. The undoubted benefits of automating the decision-making process are widely recognised. The efficacy of these procedures relies on machine learning algorithms for automated processing and analysis of a large volume of legal documents. Benefits result in increased accuracy and the ability to identify key aspects. Forecasting case outcomes reduces the time spent on preliminary case analysis. Another practical application of AI in legal scenarios related to the judge's cognitive process is the introduction of electronic systems for managing trials, including electronic filing, online access to information, and virtual court hearings. Across the globe, numerous AI tools are used to automate the search and analysis of legal information, including precedents, regulations, and court decisions. What is worth noting is that they serve all actors of the legal market by enhancing the gathering of accurate and comprehensive information. In our study, we will focus on a selected aspect of the court's cognitive process, attempting to answer the question of how certain types of automated systems support adjudication. To make these considerations more concrete, let us point to some examples of AI tools. Here, we can point to instruments from common law systems, the United States and the United Kingdom. Providing examples will later help us distinguish between the two types of mechanisms supporting the fulfilment of the judicial function (Getman, YAROSHENKO, Shapoval, PROKOPIEV, Demura, 2023).

On the one hand, we have a rules-based system, and on the other, we face a case-based methodology (Laato, Tiainen, Islam &Mäntymäki, 2022). The implementation of rule-based AI systems in the legal field streamlines the decision-making process by automating the application of legal principles to relevant facts. This approach not only enhances efficiency but also ensures consistency in the application of the law. By defining explicit criteria for decisions, rule-based AI systems provide a transparent and objective basis for legal judgments, reducing subjectivity and potential bias (Zafar, 2024; Islam, 2018).

Beyond the traditional rule-based artificial intelligence models, the legal field has seen the emergence of the case-based reasoning (CBR) model of AI. This paradigm significantly diverges from the rule-centric approaches. Unlike its rule-based counterparts, which operate within a rigid framework of pre-defined rules, the CBR model adopts a more flexible, case-by-case methodology (Feng KJK et al. 2023).

Into the first category—rules-based systems — fall COMPAS and e-Discovery tools; into the second— case-based model —system called Predictice.

One of the most widely used systems is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)[8], which has successfully developed a risk-assessment tool for predicting recidivism in the American justice system. COMPAS is a comprehensive algorithm that determines a person's likelihood of reoffending after conviction and the selection of optimal alternative sanctions or supervision of convicts. The system analyses a variety of indicators (sensitive data), including criminal history, social status and other factors[9]. However, the use of the COMPAS system in trials in the United States has sparked numerous debates regarding its ethical implications and fairness. Some critics emphasise the possibility of systemic biases and inequalities in risk assessment, depending on the racial or social characteristics of the individuals being assessed (Jackson, Mendoza, 2020; Bahl, Topaz, Obermuller, Glodstein & Sneirson, 2024).

Another popular system explored in the USA, based on rules-driven model, is e-Discovery. The process of eDiscovery involves the discovery, collection, analysis, and processing of electronic evidence, such as emails, documents, files, and other electronic data that have legal significance. In the United States, the eDiscovery system plays a crucial role in streamlining and enhancing court procedures by effectively managing vast amounts of electronic evidence. This system, therefore, supports the organisation of information and aids cognitive processes. Powered by AI and machine learning, this system can automatically identify and categorise documents, emphasise key terms, conduct information searches, apply filters, and generate comprehensive analytical reports with evidence-backed recommendations. This system serves all participants in the proceedings and is praised for it precision and uniformity in evidence management, thereby expediting the retrieval of essential information for case parties and facilitating efficient preparation for courtroom proceedings (Nogueira, 2017).

There are several AI systems in the UK that improve legal work; one of the most popular, "Predictice", could serve us as an example of case-driven models (https://predictice.com). It is the most popular platform for predicting court decisions. Designed to analyse large volumes of court documents, this platform uses machine learning algorithms to help lawyers understand potential court rulings. The UK uses the Predictice system to support lawyers' strategic decisions during legal proceedings; in fact, all participants in proceedings benefit from this tool. Based on data analysis, the platform considers factors such as prior precedents, case characteristics, and judges' information to predict possible court sentences. With this data, Predictice provides case-specific predictions. The use of the Predictice system significantly

8) https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf
9) https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

improves lawyers' understanding of potential court decisions, which, in turn, affects defence strategy and positions taken during court proceedings (Getman, YAROSHENKO, Shapoval, PROKOPIEV, Demura, 2023). However, using Predictice to predict court decisions may lead to insufficient accuracy and reliability in the results. Even with a large amount of data, algorithms may not account for complex cases without prior records (Zafar, 2024).

Based on the criterion of autonomy, we are referring to two possible models for automating judicial proceedings. First, an unassisted AI tool or system that can adjudicate legal cases unassisted (adjudicating instead of a human). The second consists of assisted AI tools (so-called a judicial decision support system (JDSS). In the second model, the system would only support the human judge by finding relevant provisions, analysing cases, and reviewing the literature. It would finally suggest a decision to the judge. Let us note, as an aside, that JDSS, i.e., the support system, seems prima facie less questionable with respect to issues such as judicial discretion. Classic concepts of judicial cognition —the process by which judicial cognition is formed —can be summarised as, in simple terms, the process by which a criminal court (judge) builds its knowledge of a case, then provides specific reasoning and formulates assessments, ultimately culminating in an evaluation of the subject matter of the criminal proceedings. The use of AI tools can interfere with cognitive processes, so it is essential to determine their limits. From the perspective of the current regulatory framework, support systems can be used much more widely (see examples in the following section) than models based on unassisted adjudication. At first glance, these supportive tools appear to carry a much lower risk of malfunction than systems that replace humans. However, it turns out that between the 'law in the books' and the 'law in action', heuristics mark an area.

In summary, the conventional method in Western legal systems for resolving legal issues involves constructing legal syllogisms. This procedure consists of three steps. First, the factual grounds for the decision must be elaborated. A legal norm must be interpreted (may it be a legal norm from positive law as in continental European systems or from case law as in common law), then in third step a factual situation is "subsumed" under the norm, i.e. all conditions provided by the norm are checked against the facts, and the respective findings of this comparison constitute the final result – judicial decision that is a concret and individualised legal norm (Duarte d'Almeida, 2019). The syllogism proceedings are juxtaposed by additional patterns anchored in heuristics. For example, a judge facing a heavy workload (related to fact-finding) will usually develop, within a short time, heuristics to cope with it. These patterns must be carefully evaluated when the assessment process is assisted by automatic tools.

On the one hand, there is an assumption that the judge applies the law by making fully rational choices. The assumption of rationality entails accepting several idealised assumptions about the knowledge and behaviour of a rational actor. In practice, however, these assumptions do not always hold true—for example, the assumption of the completeness of knowledge, the assumption of the optimisation of working time, and the reliability of action. In other words, lawmakers assume that those applying their laws apply rational choice theory, i.e., that actors will weigh the consequences provided by legal norms and, based on logical deduction, choose the one that maximises their advantages or utility. Given that lawmakers do not intend only to provide a set of rules for equal justice for all actors bound by them, these aims would be jeopardised if the behaviour of those bound by these laws does not follow the patterns assumed by the lawmaker (compare Martinho, 2025, p. 569). Here is a space for heuristics that always carry the risk of bias. Regarding the potential use of AI to automate adjudication heuristics that simplify work, this should be mentioned. Here, one shall refer to research on the psychological issue of the «persuasiveness» of judges' supporting systems: on the one hand, fully automated, and on the other, supportive. One of research said that *"while it may seem logical to draw a distinction between fully automated decision-making and semi-automated decision making, in practice the boundaries between the two are blurred and (g)iven the pressure of high caseloads and insufficient resources from which most judiciaries suffer, there is a danger that support systems based on artificial intelligence are inappropriately used by judges to «delegate» decisions to technological systems that were not developed for that purpose and are perceived as being more «objective» even when this is not the case. Great care should therefore be taken to assess what such systems can deliver and under what conditions, may be used in order not to jeopardise the right to a fair trial"* (Djikstra, 2001, p. 119).

Although legally trained, judges are also generally subject to biases such as those described above; other biases are even more relevant when making legal decisions, such as hindsight bias and the "anchor" bias (Hoffman, 2020). Both phenomena affect the judge's processes, regardless of whether the judge uses additional AI tools for adjudication. However, when these auxiliary tools are used, both heuristics might be strongly reinforced.

Hindsight bias describes the phenomenon that, given the factual outcome of a development, the probability of that outcome seems much higher than that of others when viewed ex post. At court, this bias may result in defendants being judged as capable of preventing a bad outcome; for example, in cases where there is an assumption of risk, hindsight bias may lead judges to perceive the event as even riskier merely because of the poor outcome. This may lead the judge/jury to feel that the plaintiff should have exercised greater caution in the situation, although, from an a priori perspective, this caution would not have been deemed necessary (Hoffamn, 2020). In the context of AI-assisted adjudication, hindsight bias may be reinforced because the court's cognitive process is shortened and part of its own analysis is replaced (or enriched) by AI-generated auxiliary data.

However, the most dangerous aspect of automatic decision-support systems is heuristic bias, which includes anchoring and adjustment. In practice, no less critical bias describes the tendency to rely too heavily on the first piece of information offered (the "anchor") when making decisions. For example, suppose the AI system suggests a certain period of imprisonment for the accused as an average for similar cases. In that case, the judge might be automatically influenced by this "anchor" when issuing a final decision. The same applies to all AI tools that interpret data to support the judge's cognitive process. There is no doubt, therefore, that technological improvements in decision-making processes must be accompanied by continuous monitoring of the systems used, as well as user self-monitoring, to prevent the recurrence of negative patterns (Hoffman, 2020).

*

Most importantly, the use of AI decision support systems must be fully transparent and controllable from the perspective of all participants in the proceedings. As indicated in the first part of the discussion, the undoubted benefits of using AI tools are accompanied by serious concerns about the violation of fundamental rights, including the right to defense. The right to court presupposes the observance of the standard of a fair trial and the issuance of a judgment by an independent and impartial court. In this context, it is worth reflecting that AI tools cannot undermine the court's standard of autonomy and independence, which is granted by the cognitive processes' autonomy of a judge. As indicated earlier, it seems that criminal law systems can use AI tools, even in the form of an e-court, but only in minor cases and when a preferential procedure for the accused and an adequate control mechanism are ensured. All instruments supporting the adjudication process must meet the conditions of full transparency and all standards elaborated in ethical Aquis, based on legal, ethical, and pragmatic norms (see the concept of Trustworthy AI). The use of auxiliary AI tools for adjudication in a semi-formal manner, without the need for detailed documentation of their use, seems to be the most challenging.

Bibliography

Bahl U.; Topaz C.; Obermüller L.; Goldstein S.; Sneirson M. (2025), Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida" UCLA Law Review. Retrieved March 10, 2025.

Caroline Gans-Combe (2022), Automated Justice: Issues, Benefits and Risks in the Use of Artificial Intelligence and Its Algorithms in Access to Justice and Law Enforcement, Ethics, Integrity and Policymaking, vol 9 .

Djikstra (2001), Legal Knowledge-based Systems: The Blind leading the Sheep?, International Review of Law, Computers & Technology (2001), Vol. 15, No. 2, pp. 119– 128.

Duarte d'Almeida, Luís (2019), On the Legal Syllogism, in David Plunkett, Scott J. Shapiro, and Kevin Toh (eds), Dimensions of Normativity: New Essays on Metaethics and Jurisprudence (New York, online edn, Oxford Academic, February 21 2019), https://doi.org/10.1093/oso/9780190640408.003.0015.

Fernández-Martínez C, Fernández A (2020), AI and recruiting software: ethical and legal implications. Paladyn. 11(1):P199. https://doi.org/10.1515/pjbr-2020-0030.

Feng KJK et al. (2023), Case repositories: towards case-based reasoning for AI alignment arXiv (Cornell University) https://doi.org/10.48550/arxiv.2311.10934.

Floridi L. (2018), Soft Ethics and the Governance of the Digital, Philosophy & Technology, March 2018, Volume 31, Issue 1.

Fiori L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018); AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach (Dordr). 2018;28(4):689-707. doi: 10.1007/s11023-018-9482-5. Epub 2018 November 26. PMID: 30930541; PMCID: PMC6404626.

Getman A.P., OLEG M. YAROSHENKO, Roman V. Shapoval, Roman Ye. PROKOPIEV, Maryna Demura 2023. THE IMPACT OF ARTIFICIAL INTELLIGENCE ON LEGAL DECISION MAKING, International Comparative Jurisprudence 2:155-169. https://www.ceeol.com/search/article-detail?id=1210529

Hoffmann T. (2020), Heuristics in Legal Decision-Making. Acta Baltica Historiae et Philosophiae Scientiarum 1:62-71.

Islam MB, Governatori G. (2028), RuleRS: a rule-based architecture for decision support systems, Artificial Intelligence Law. ;26:315.

Jackson E, Mendoza C. (2020), Setting the record straight: what the COMPAS core risk and need assessment is and is not. Harvard Data Sci Rev.;

Janusz-Pohl, B. (2022). The model of the Polish criminal procedure. An analysis from the perspective of its structure and the form of process functions and procedural guarantees. W B. Janusz-Pohl, Ł. Pohl, & W. Achrem (Redaktorzy), Forensic identification based on biological testing in cross-border criminal proceedings: legal, methodological and praxeological aspects. [Vol. 1]. Wydawnictwo Naukowe Uniwersytetu Szczecińskiego. https://zachodniopomorska.policja.gov.pl/sz/aktualnosci/projekty-realizowa/interreg-110/22509,Korelacja-identyfikacji-i-zwalczania-transgranicznych-powiazan-terrorystycznych.html

Janusz-Pohl, B. (2023). Theoretical and methodological foundations for consensual models based on Polish example. W S. Pawelec (ed.) Consensual Mechanisms in Criminal Proceedings – Integrative and Comparative Perspective. Peter Lang Publishing Group. https://www.peterlang.com/document/1365342

Janusz-Pohl, B. (A2025). The language of defence in criminal proceedings: The Polish perspective. W J. Visconti (ed.), The Language of Lawyers. A European Perspective. De Gruyter. https://doi.org/10.1515/9783111340982-029

Janusz-Pohl, B. (B2025); Perceptions of the Effectiveness and Its Interdisciplinary Approach – Keynotes. W D. Vicoli (ed.), Effective Justice. International and Comparative Approaches. Volume 1 (T. 38). Peter Lang Publishing Group.

Janusz-Pohl, B. Theoretical and Methodological Foundations for Consensual Models Based on Polish Example". Consensual Mechanisms in Criminal Proceedings – Integrative and Comparative Perspective,

Szymon Pawelec (ed.), Peter Lang Publishing Group, 2023, https://www.peterlang.com/document/1365342.
Laato S, Tiainen M, Islam AKMN, Mäntymäki M. (2022), How to explain AI systems to end users: a systematic literature review and research agenda. Internet Res;32(7):1. https://doi.org/10.1108/INTR-08-2021-0600.
Martinho, A. (2025), Surveying Judges about artificial intelligence: profession, judicial adjudication, and legal principles. AI & Soc 40, 569–584 https://doi.org/10.1007/s00146-024-01869-4
Nogueira MG, et al (2017), E-discovery as a mean to improve information security. In: presented at the 2017 Computing Conference https://doi.org/10.1109/SAI.2017.8252214.A.
Said G, Azamat K, Ravshan S, Bokhadir A.(2023), Adapting legal systems to the development of artificial intelligence: solving the global problem of AI in judicial processes. Int J Cyber Law. https://doi.org/10.59022/ijcl.49.
Zafar, A. (2024), Balancing the scale: navigating ethical and practical challenges of artificial intelligence (AI) integration in legal practices. Discov Artif Intell 4, 27 https://doi.org/10.1007/s44163-024-00121-8
Zhang, Q., Chen, S., Song, T. (2024), A Literature Analysis of the Application of Artificial Intelligence in Judicial Adjudication. In: Zhao, F., Miao, D. (eds) AI-generated Content. AIGC 2023. Communications in Computer and Information Science, vol 1946. Springer, Singapore. https://doi.org/10.1007/978-981-99-7587-7_21
Ziembiński, Z. (2021), Chapter 7 Norms of Conduct, In Poznań School of Legal Theory. Leiden, Niederlande: Brill. https://doi.org/10.1163/9789004448445_8

인공지능 시대의 공공성 재고 – 드러남의 공간을 중심으로

Rethinking Publicness in the Age of AI: Focusing on the Space of Appearance

허유선
경남대학교 교수

Eusun Heo
Professor, Kyungnam University

초록

오늘날 자동화된 알고리즘에 기반한 디지털 플랫폼은 인간의 관계 맺기와 공간 경험을 근본적으로 변화시키고 있다. 새로운 기술이 제시하는 공간은 보편적 접근성을 약속하는 듯 보이나, 실제로는 새로운 불평등과 배제의 장으로 작동한다. 본 논문은 자동화된 알고리즘이 구성하는 공간이 공공성의 근본 조건인 '드러남의 공간(space of appearance)'과 긴장 관계에 놓임을 보인다. 아렌트의 공공성 개념에 따르면, 공적인 영역은 타자에게 자신을 드러내고 응답받을 수 있는 가능성 위에 성립한다. 그러나 알고리즘의 선제성·비가시성·조건부 재현은 이러한 가능성을 제한할 수 있다. 따라서 자동화된 알고리즘 시대의 공공성에 대한 지속적 성찰과 대안 모색이 요청된다.

Abstract

Today, digital platforms driven by automated algorithms are fundamentally transforming human relationships and spatial experience. The spaces created by new technologies appear to promise universal accessibility, yet in practice, they often operate as new arenas of inequality and exclusion. This paper demonstrates that algorithmically constructed spaces stand in tension with the space of appearance, the fundamental condition of publicness. According to Arendt's conception of publicness, the public realm is grounded in the possibility of revealing oneself to others and receiving their responses. However, the preemption, invisibility, and conditional representation of algorithms can restrict such possibilities. Therefore, continuous reflection and the search for alternatives are required to rethink publicness in the age of automated algorithms.

1. 문제 제기: 자동화된 알고리즘과 새로운 공간, 공공성

오늘날 자동화된 알고리즘에 기반한 디지털 플랫폼은 인간의 관계 맺기와 그 장場으로서 공간 경험을 근본적으로 변화시키고 있다. 화상 수업, 원격 회의, 전자 상거래, 메타버스 등이 그 대표적 사례이다. 이들은 기존의 공간적 제약을 넘어선 새로운 공간을 제시하며, 이는 보편적 접근성과 상호작용성을 약속하는 듯 보인다. 그러나 자동화된 알고리즘이 만드는 공간은 새로운 배제의 공간이기도 하다.

첫째, 알고리즘은 기존의 사회적 불평등을 공간적 차원에서 재현한다. 승차 호출 앱의 요금 산정 시스템은 동일한 거리와 조건에서도 지역의 인구 구성이나 소득 수준에 따라 가격을 다르게 책정하며, 특정 지역을 회피하거나 접근성을 낮추는 방향으로 작동한다(Pandey & Caliskan, 2021). 메타의 주택 광고 알고리즘 또한 인종, 성별, 장애 여부 등의 특성에 따라 광고 노출을 조정함으로써, 거주지 선택과 이동 가능성 등을 차등적으로 분배했다(Department of Justice, 2022). 이러한 사례들은 인공지능 윤리 논의에서 인공지능의 편향성과 차별 문제라는 범주로 다루어져왔다.

둘째, 자동화된 알고리즘은 공간 형성의 단계부터 개입하여, 해당 공간에 드러나 공동체의 논의에 참여할 수 있는 존재를 규정한다. 특정한 자원 분배의 차별을 넘어서, 그 분배에 참여할 수 있는 자격 조건 자체를 제한하는 것이다. 예를 들어, Shim(2023)은 '스마트도시'가 정보 제공에 참여하지 않는 존재를 비가시화한다고 지적한다. 스마트도시는 디지털 공간과 물리적 공간이 통합되는 일종의 메타적 플랫폼으로서 우리의 상호작용을 매개하며, 도시 거주자들이 지속적으로 정보를 제공하고 기술적 조정에 '참여'할 것을 요구한다. 스마트도시는 오직 알고리즘이 요청하는 '참여'를 '수행'하는 존재만을 가시화한다. 기술적 참여에 부합하지 않는 존재는 일종의 유령적 존재로 간주되는 것이다.
다시 말해 자동화된 알고리즘은 새로운 사회적 상호작용의 장을 구성하는 동시에, 상호작용을 위해 타자에게 드러날 기회에 개입하며, 구조적 배제 장치로 작동한다. 이는 누가 타자에게 보이고, 들리고, 말하고, 응답받을 수 있는 존재로 인정되는가, 곧 누가 공적인 것으로 드러나고 공적인 것에 참여할 수 있는가에 관한 문제이기도 하다. 아렌트에 따르면, 타자에게 보여지고 들려질 수 있는 '드러남의 공간(space of appearance)'은 공적인 것을 형성하는 일에 참여할 수 있는 가능성으로서, 공공성의 필수 조건이다(Arendt, 2002). 그러므로 자동화된 알고리즘에 기반한 새로운 공간은 우리시대 공공성에 대한 재고를 요청한다.

2. '드러남의 공간'으로서 공공성

공공성(Öffentlichkeit)은 크게 세 가지 의미, 곧 국가와 관련된 공적인(official) 것, 특정 개인 및 집단의 이해관심에 국한되지 않고 모든 사람과 관계된 공통적인(common) 것, 비밀이나 프라이버시와 대조적으로, 누구에게나 열려 있는(open) 것이라는 의미로 논의된다. 이 세 가지 의미 중 어느 것에 초점을 맞추든, 공공성은 '사적인 것(the private)'과 구분되는 것으로서 공적인 것(the public)을 전제할 때 가능하다. 그러나 이 구분은 고정된 것이 아니라, '공적인 것'의 담론 형성 과정에 의존하는 유동적인 것이다(Saito, 2024). 따라서 공공성은 고정된 개념이라기보다, 공적인 것의 논의와 참여에 따라 그 의미가 변화하는 구성적 개념이라 할 수 있다. 공공성 개념은 공적인 것과 사적인 것의 구분에 의존하며, 누가 공적인 것에 참여할 수 있는지의 문제와 밀접하게 관련된다. 아리스토텔레스의 가정과 폴리스의 구분, 하버마스의 부르주아 공론장은 공공성이 보편적 개방성과 접근성을 표방하면서도 특정 계층이나 조건을 전제로 형성되었음을 보여준다. 곧, 공공성 개념의 변천은 재산, 교육, 젠더 등의 조건에 따라 누가 공적인 것에 참여할 수 있는지를 선별해 온 배제의 구조를 담고 있다.

아렌트의 공적인 것에 관한 논의는 바로 이 배제, 공적인 것으로 접근 및 참여의 박탈 가능성에 주목하며, 그 현존이 타자에게 드러날 수 있는 '드러남의 공간'을 공공성의 근본 토대로 제시한다. 아렌트는 '사적'이란 용어가 원래 '박탈된'이라는 의미를 가졌음을 부각한다. 완전히 사적인 생활을 한다는 것은 우선 진정한 인간에게 필수적인 것이 박탈되었음을 의미한다(Arendt, 2019, p. 142). 그렇다면 진정한 인간의 삶을 위해 필수적인 것으로서 공적인 것이란 무엇인가?

첫째, 공적인 것은 '드러남의 공간'이다. 이 공간에서 나는 타인에게, 타인은 나에게 나타난다(Arendt, 2019, p. 300). 나타난다는 것은 타자에게 나타남 곧, 나뿐만 아니라 다른 사람도 보고 들을 수 있는 것이라는 의미로, 가장 넓은 의미의 공개성을 뜻한다. 이러한 현상의 공간은 사람들이 함께 사는 곳이면 어디나 존재하고, 따라서 공적 영역의 모든 형식적 구조와 다양한 형태의 정부에 앞서 존재한다(Arendt, 2019, p. 301). 그러나 이 공간은 단순한 물리적 연장(extension)을 의미하는 것이 아니라, 사람들의 행위(action)에 의존하는 잠재적 공간으로, 행동하고 말하는 사람들 '사이(in-between)'에 나타나는 공간이다.

아렌트에게 '행위'란 '노동(labor)', '작업(work)'과 구분되는 인간의 활동으로, 사물이나 물질의 매개 없이 인간들 사이에서 직접적으로 이루어지는 것이다(Arendt, 2019, p. 83). 이는 말과 행동을 모두 포함한다. 말과 행동을 통해 인간은 자신과 타인을 구별하고, 자신을 세계에 전달한다. 그러므로 행위의 근본 조건은 다수의 인간이 이 지구상에 살고 세계에 거주한다는 사실에 상응한다(Arendt, 2019, p. 84). 그런데 인간의 다원성은 동등과 차이라는 이중의 성격을 지닌다(Arendt, 2019, p. 273). 사람이 동등하지 않다면 말과 행동이 불가능할 것이며, 사람들이 다르지 않다면 타인에게 자신을 이해시키기 위한 말과 행동이 필요하지 않기 때문이다.

말과 행동은 인간이 단순히 물리적 대상으로서 육체적 존재를 드러내는 것이 아니라, 타자와의 동등성과 차이를 지닌 고유한 인격적 존재로서 서로에게 자신을 드러내는 방법이다. 신체적 드러남은 별도의 활동 없이도 신체 자체의 형태와 목소리를 통해 나타난다. 그러나 자신이 '누구'인지를 뜻하는 인격적 정체성은 오직 말과 행동을 통해서만 드러난다. 이러한 말과 행동의 계시적 성질은 오직 타인과 함께 존재하는 곳에서만 곧, '함께함'에서만 나타난다(Arendt, 2019, p. 278). 그러므로 공적인 영역이란 (인격으로서) 인간이 스스로를 드러내기 위해 반드시 필요로 하는 공간이라 할 수 있다(Arendt, 2019, p. 310). 그러나 이는 하나의 고정된 정체성이 선재하고, 이것이 말과 행동을 통해 외화되어 드러난다는 의미가 아니다. '누구'라는 정체성은 행동이나 말에 대한 타자의 응답으로 비로소 생성되며(Saito, 2024, p.62), 곧 타자의 존재와 함께 생성되는 것이다.

타자와 함께 잠재적인 공간을 형성하는 것으로서 말과 행위의 특성은 공적인 것의 두 번째 의미로 나아간다. 공적인 것은 사적 소유지와 구별되는 '공동 세계' 자체를 뜻한다. 그러나 이는 우리가 '세계'라는 표현에서 일반적으로 떠올리는 자연이나 지구와는 달리, 인간이 제작한 인공물과 인위적 세계에 거주하는 '사람들 사이에서' 일어나는 사건과 관련된다. 곧, 사물의 세계를 공동으로 취하는-마치 탁자가 그 둘레에 앉는 사람들 사이에 자리잡고 있듯이- '사이' 공간에서 함께 살아있음을 의미한다(Arendt, 2019, p. 135).

그러므로 아렌트에게 공적인 것의 핵심은 모든 형식에 앞서 있는, 환원불가능한 고유한 존재로서 타인 앞에 자신이 드러날 수 있고, 타인의 존재를 마주할 수 있으며, 그로부터 사이 공간으로서 공동 세계를 형성할 수

있는 가능성이다. 이러한 점에서, 아렌트의 논의는 추상적인 공공성 논의를 넘어, 물리적 공간과 담론의 공간을 모두 포괄하는 구체적이고 현실적인 공간 논의와 직접적으로 연결된다.

3. 자동화된 알고리즘과 '드러남의 공간'

문제는 자동화된 알고리즘, 곧 오늘날 '인공지능'이라 부르는 것이 공공성의 조건으로서 '드러남의 공간'과 긴장 관계에 놓인다는 것이다.

기존 자동화 기술과 차별점으로서 자동화된 알고리즘의 특성은 통계 기반의 선제적(pre-emptive) 예측과 비가시성에 있다(Andrejevic, 2021). 알고리즘의 선제적 예측 혹은 선점(pre-emption)은 타자와의 조우가 일어나기 '이전' 단계에서 결과를 예측하고, 예측된 결과를 성취하거나 피하도록 자원을 할당함으로써 사람들의 행동을 '미리' 조정하여, 아렌트가 말한 드러남의 공간에 참여할 참여 가능성을 사전에 선별·차단하는 장치로 작동한다. 예를 들어, 추천 알고리즘을 활용한 맞춤형 콘텐츠 소개 혹은 관심 없는 콘텐츠의 배제 등은 사람들이 자각하기도 전에, 인지의 범위, 이질적인 타자와의 마주침, 상호작용의 가능성을 조정한다. 이는 개인의 행동을 조율할 뿐만 아니라, 우리가 공간을 인식하고 경험하는 방식 및 구성 방식, 곧 공간에 대한 관계적 감각에도 영향을 미친다. 가령, 교통 경로 안내 시스템은 일견 효율성과 안전을 제공하는 것처럼 보이지만 특정 지역이나 커뮤니티를 지속적으로 회피하거나 제외하는 방식으로 동작할 수 있으며, 이로써 사회적 낙인을 강화하고 공간적 고립을 심화시킬 수 있다.

자동화된 알고리즘의 또 다른 특징은 개인에게 직접적이고 가시적인 강제로 작동하는 대신, 인구 집단에 영향을 주는 다층적 요소가 결합된 환경을 조정하는 비가시적인 방식으로 작동한다는 것이다. 예를 들어, 송도 스마트도시의 '스마트 빌리지'는 입주민이 주거 혜택을 받는 대신, 개인정보를 제공하며 기술 실증 실험에 참여해야 한다(Ministry of Land, Infrastructure and Transport, 2020). 거주자의 활동은 도시 환경을 이루는 수많은 센서와 시스템을 통해 실시간으로 감시된다.

스마트도시에서 모든 거주자의 말과 행동은 필수적으로 드러나야 하는 것이 되지만, 그 드러남은 다양한 타자와의 관계 속에서 형성되는 개인의 고유성을 표현하는 방식과는 거리가 있다. 자동화된 알고리즘은 각 개인을 고유한 정체성을 지닌 주체로서가 아니라, 데이터 항목의 조합으로 환원하고, 통계적 인구 집단 속 하나의 변수로 다룬다. 곧, 스마트도시에서 드러남이 허용되는 주체는 고유한 자아가 아니라, 알고리즘의 모델에 부합하는 알고리즘적 재현이다. 이는 고유한 존재의 자유로운 행위를 제약할 뿐만 아니라, 다양성의 기반마저 약화시킬 수 있다.

한편 스마트도시 내 시민 참여의 필수성은 일상화되고 자동화된 감시의 또 다른 이름이기도 하다. 결과적으로, 도시 공간 전체가 감시와 효율에 따라 구성되는 공간으로 전환되는 것이다. 알고리즘은 특정 개인이나 집단을 직접적으로 억압하거나 배제하는 방식이 아니라, 일상적이고 중립적으로 보이는 환경 구성 속에 개입한다. 자동화된 감시와 조정은 그 작동 방식이 뚜렷하게 인식되지 않기 때문에 대응이 더욱 어렵다.

이는 애초에 알고리즘이 불투명하다는 사실 곧, 알고리즘의 불투명성(opacity)을 환기시킨다. 인간의 학습 방식과 다른 알고리즘의 기계학습은 인간이 인공지능 알고리즘의 작동을 온전히 이해하거나 파악하는 것을 어렵게 만든다. 입력값(input)과 출력값(output)을 볼 수는 있지만, 어떤 과정을 거쳐 어째서 이런 결과가 도출되었는지를 투명하게 파악하고, 이해하기는 어렵다. 알고리즘의 자동화된 의사결정 절차가 인간에게는 일종의 '블랙박스'처럼 여겨질 수 있는 것이다. 이러한 기술적 불투명성은 기업의 영업비밀, 정보 비공개 정책 등 현실적인 불투명성 조건과 맞물리면서(Ko, Jeong, & Park, 2019), 일반 시민이 자동화된 알고리즘의 영향을 인식하거나 문제 제기를 하는 것을 어렵게 만든다. 이처럼 비가시적이고 불투명한 시스템은 자동화된 결정의 효과와 영향력을 외부로부터 감추며, 사회적 책임의 주체를 식별하는 것 또한 어렵게 만든다. 이에 상응하여, 공적인 것에 참여할 수 있는 기회의 구조적 배제를 인지하거나 저항하기는 더욱 어려워진다.

결론적으로, 자동화된 알고리즘은 공간의 구성과 주체의 드러남에 깊숙이 관여하면서, 고유한 현존의 자유로운 드러남을 통계적 예측과 선제적 조정이 매개하는 선택적 드러남으로 대체한다. '누가', '어떻게' 드러날 수 있는지가 기술적 구조와 함께 사전에 규정되고, 조건적으로 허용되는 것이다. 이는 자동화된 알고리즘이 매개하는 공간이 단지 기술의 유익한 적용 문제를 넘어서, 공적인 것으로 드러나고 공적인 담론에 참여할 수 있는 공공성의 문제임을 시사한다. 그러므로 자동화된 알고리즘의 작동 기제에 대한 비판적 성찰과 이에 대응하는 우리의 실천은 오늘날 공공성이 어떻게 구성되는지, 그리고 공공성을 어떻게 실현할 것인지에 대한 물음으로 이어진다.

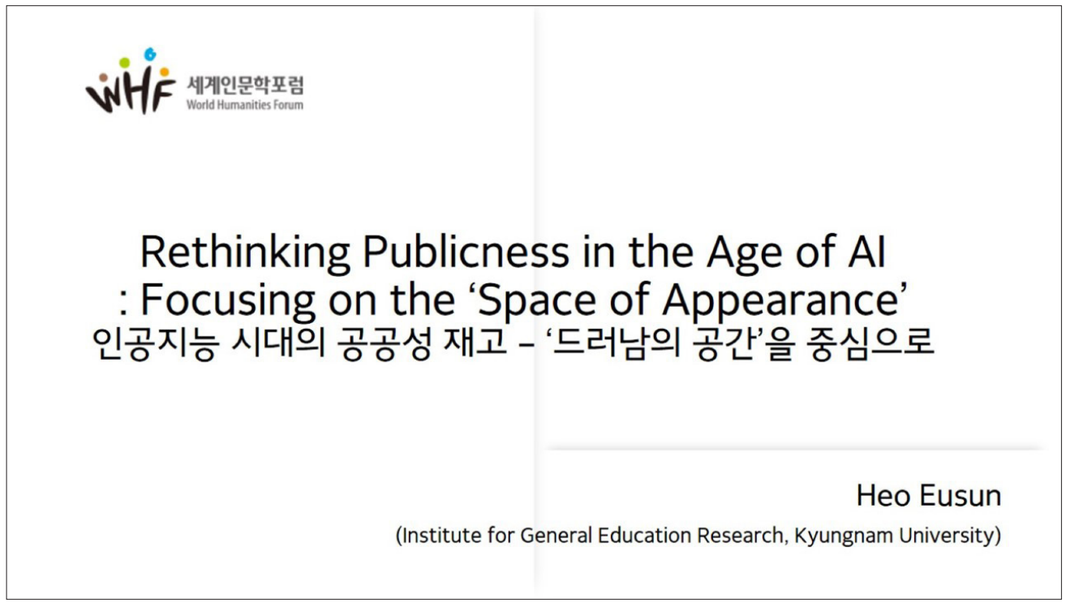
4. 인공지능 시대의 공공성 논의, 향후 과제

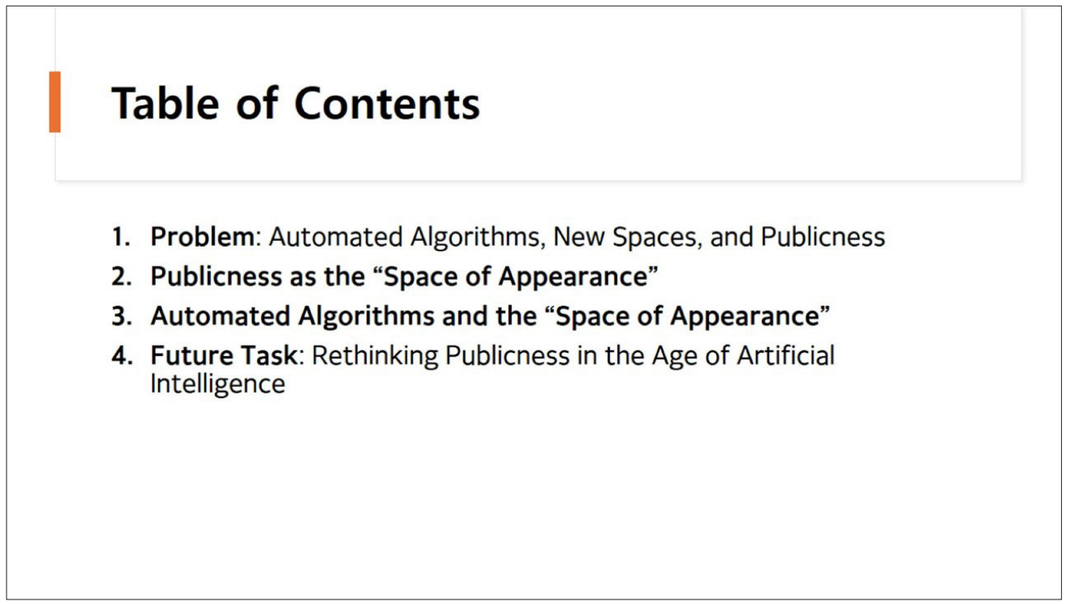
이로부터 다음의 후속 논의가 요청된다. 첫째, 자동화된 알고리즘에 기반한 공간을 다양한 타자들의 고유성이 드러나고 상호교류가 발생하는 공적인 공간으로 재전유·구성하는 실천적 방안의 모색이 필요하다. 이를 위해서는 기존의 기술·경제적 효율성 중심 기획을 넘어, 공간 설계의 초기 단계부터 공공성을 적극적으로 고려하는 관점이 요구된다. 예를 들어, EU의 스마트도시전략은 인간의 보다 나은 삶이 지속가능한 공동체의 구현에 초점을 맞춘다(Manville et al., 2014, p. 9). 덴마크 올보르(Aalborg)는 장애인의 접근성을 높이기 위해 시민 주도의 문화 활동 지원 서비스를 개발했으며, 영국 버밍엄(Birmingham)은 장애 여성의 이동권을 보장하기 위한 프로그램을 운영하고 있다(Oliveira & Campolargo, 2015, p. 2342). 더불어 공간 기획 및 활용 과정에서 시민 참여가 더욱 확대되어야 한다. 공간으로의 드러남이나 참여가 단순한 데이터 제공이나 서비스 이용으로 환원되지 않고, 정책·설계·감사 단계마다 다양한 의사소통의 장이 마련되어야 하며, 이러한 논의가 실제 도시 공동의 의사결정에 반영되어야 한다.

그러나 자동화된 알고리즘의 선제성·비가시성·조건부 재현 및 환원이라는 특성을 고려할 때, 이들이 형성하는 공간은 공공성의 근본 조건과 지속적인 긴장 관계에 놓인다. 따라서 자동화된 알고리즘이 매개하고 형성하는 새로운 공간이 어떠한 존재를 드러나게 하고, 누구를 배제하는지에 대한 지속적인 성찰과 대응이 요청된다. 이는 우리 시대의 공공성을 새롭게 사유하고 실현하기 위한 과제이기도 하다.

둘째, 타자에로 드러남이라는 공공성의 근본 조건과 프라이버시의 관계가 보다 상세히 고찰될 필요가 있다. 자동화된 알고리즘 환경에서 드러남은 제한적이며, 구조적 배제를 동반한다. 그러나 공적인 것으로의 드러남은 무제한적 공개를 의미하지 않는다. 공공성은 언제나 사적 삶의 보장과 짝을 이루는 관계 속에서 성립하기 때문이다. 오늘날 프라이버시는 정보주권, 곧 자기정보통제권의 차원에서 이해되며, 개인이 언제·어떻게·얼마나 자신을 드러낼 것인지 스스로 결정할 수 있는 권리가 전제되어야 한다. 따라서 드러남의 권리와 프라이버시는 공적 관계 형성과 고유한 주체 형성의 조건으로 함께 고려되어야 한다.

마지막으로, 자동화된 알고리즘 기반 공간과 공공성 실현에 관한 논의는 보다 확장된 공공성 관점에서 다루어질 필요가 있다. 이 글은 공공성을 한나 아렌트의 '드러남의 공간' 개념에 주목하여 논의를 전개했지만, 공공성의 조건에 관해서는 다양한 접근이 가능하다. 특히 공공성이 구성적 개념임을 긍정할 때, 공공성은 추상적 원리라기보다 현장 속 다양한 실천 주체와 행위자들이 만들어가는 구체적 관계와 맥락(Cho, 2023)과 더불어 이해되어야 한다.





참고문헌

Andrejevic, Mark. (2021). Automated media (Lee, Hee Eun, Trans.). Seoul: Culture Look.

Arendt, H. (2019). The human condition. Seoul: Han-gil Great Books.

Cho, Mun Young. (2023). Becoming Public: The Assemblage of the Public Housing Project near Seoul Station. Cross Cultural Studies, 29(2), 279–332. https://doi.org/10.17249/CCS.2023.12.30.2.279

Department of Justice. (2022, June 21). Justice Department secures groundbreaking settlement agreement with Meta Platforms, formerly known as Facebook, to resolve allegations of discriminatory advertising. U.S. Department of Justice. https://www.justice.gov/archives/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known

Ko, Haksoo, Jeong, Hae Bin, & Park, Dohyun. (2019). Artificial intelligence and discrimination. The Justice, (171), 199–277. https://doi.org/10.29305/tj.2019.04.171.199

Manville, C., Cochrane, G., Cave, J., Millard, J., Pederson, J. K., Thaarup, R. K., Liebe, A., Wissner, M., Massink, R., & Kotterink, B. (2014). Mapping smart cities in the EU (Study report No. PE 507.480). European Union, Directorate General for Internal Policies. https://data.europa.eu/doi/10.2861/3408

Ministry of Land, Infrastructure and Transport. (2020, November 11). Busan Eco Delta smart city, smart village opens applications for residents [Press release]. Ministry of Land, Infrastructure and Transport. https://www.molit.go.kr/USR/NEWS/m_71/dtl.jsp?id=95084726

Oliveira, Álvaro, & Campolargo, Margarida. (2015, January). From smart cities to human smart cities. Paper presented at the 48th Hawaii International Conference on System Sciences, Kauai, Hawaii.

Pandey, A., & Caliskan, A. (2021). Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 822–833. https://doi.org/10.1145/3461702.3462561

Saito, Junichi. (2024). Publicness. Seoul: Ieum.

Shim, Hanbyul. (2023). A mundane smart city: Re-cognizing the smart city as a pervasively digitalized world. Space and Environment, 33(2), 275–320. http://doi.org/10.19097/kaser.2023.33.2.275

I. Problem
Automated Algorithms, New Spaces, and Publicness

문제 제기: 자동화된 알고리즘, 새로운 공간, 그리고 공공성

Publicness in the Age of AI: Why Rethink It?

- Automated algorithms transform **human relations (인간 관계)** and **spatial experience (공간 경험)**.
- Optimistic vision: universal accessibility and interactivity.
- Yet algorithms reproduce **new forms of exclusion (배제의 새로운 형태)**.

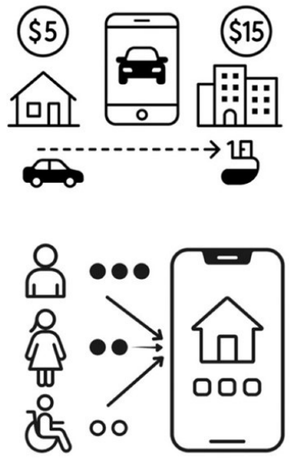
Digital Spaces as New Zones of Exclusion

- 디지털 공간, 새로운 배제의 장

1. They reproduce existing social inequalities in spatial form.
   (공간적 차원에서 재현되는 사회적 불평등)
2. The promise of openness turns into conditional visibility and participation.
   (조건적 가시성과 참여)

Algorithmic Bias and Spatial Discrimination
승차호출·주택 광고 알고리즘의 편향

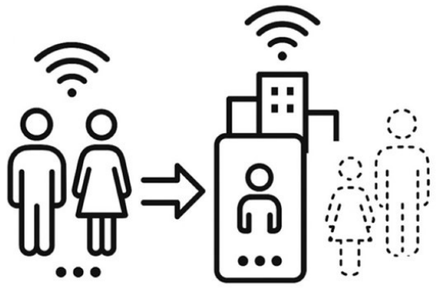
- **Ride-hailing pricing system:**
  - Fares differ by neighborhood demographics and income level — even for the same distance (Pandey & Caliskan, 2021).
    → Algorithms encourage **avoidance of low-income areas (저소득 지역 회피)** and reduce accessibility.
- **Housing ad algorithm (Meta):**
  - Exposure adjusted by race, gender, and disability (Department of Justice, 2022).
    → Unequal **distribution of housing and mobility opportunities (주거·이동 기회 분배 불평등)**.



Algorithmic Participation and Invisibility in Smart Cities

스마트도시의 참여 조건과 비가시성

- Automated algorithms intervene **from the stage of spatial construction (공간 구성 단계부터 개입)**.
- They not only reproduce inequality but also **restrict who is qualified to participate (참여 자격 자체를 제한)**.
- In Songdo's *Smart Village*, residents exchange housing benefits for personal data and participation in technological experiments.
- Only those who **perform the participation requested by algorithms (알고리즘이 요청하는 '참여'를 수행)** remain visible (Shim, 2023).
- Those outside this frame become **invisible entities, excluded from public appearance (조건에 해당하지 않으면 비가시화됨)**.



II. Publicness as the "Space of Appearance"

'드러남의 공간'으로서의 공공성

Automated Algorithms Restrict Public Appearance
자동화된 알고리즘, 공적인 것으로 드러남의 가능성을 제한

- Automated algorithms construct spaces with **dual characteristics (양가성)** : opening **new arenas of interaction** while operating as **structural devices of exclusion (구조적 배제 장치)**.
- They limit **who can be seen, heard, and responded to (타자에게 보이고, 들리고, 응답받을 수 있는 공간 제한)** within these spaces.
- Hannah Arendt, **the *space of appearance (드러남의 공간)*** - where individuals can appear before others
  - is a **necessary condition of publicness (공공성의 필수 조건)**.
- In the age of automated algorithms, **who is allowed to appear as public**, and **how can we reimagine publicness and participation?**

(알고리즘 시대, 공공성의 의미와 실현 가능성 재사유 필요)

What is Publicness(공공성, Öffentlichkeit)?

- Three Meanings of Publicness (Saito, 2024)
  - **Official:** Relating to the state or government
  - **Common:** Shared by all
  - **Open:** **Accessible** to everyone

→**Publicness** is formed in a relationship with the **"private"**(사적 영역과의 관계 속에서 형성)

- Therefore, publicness is not a fixed but a **constitutive concept**
  (공공성은 고정된 개념이 아니라 구성적 개념)

Publicness: Also a History of Exclusion (공공성은 배제의 역사이기도)

- **Aristotle:** *Oikos* vs. *Polis* (가정과 폴리스의 구분)
- **Habermas:** *The Bourgeois Public Sphere* (부르주아 공론장)

- Publicness (공공성) claims **universal openness (보편적 개방성)**, yet it has always relied on **specific classes and conditions (특정 계층과 조건)**.
- It carries a **structure of exclusion (배제의 구조)**, determining participation by **property, education, and gender (재산, 교육, 젠더 등)**.

- Hence the enduring question: **"Who is allowed to participate in the public realm?" (누가 공적 영역에 참여할 수 있는가)**

Arendt's, The "Space of Appearance" (한나 아렌트, 드러남의 공간)

- The space where one can be **seen, heard, and responded to by others** (타자에게 보이고, 들리고, 응답받을 수 있는 공간)
- Exists wherever people live, **preceding all forms and institutions** (사람이 함께 살아감으로써 어디서든 존재하며, 모든 형식과 제도에 앞선다)
- **Not a fixed location**, but arises **"in-between" (사이)** where **speech and action** occur (물리적으로 고정된 장소 아니라, 말과 행동이 오가는 '사이'에서 성립)
- Through speech and action, **diverse individuals form a common world** (서로 다른 사람들이 말과 행동을 통해 공동 세계를 형성)

→ **Publicness (공공성)** is founded upon this *space of appearance*, the **essential condition of human life (인간적 삶의 필수 조건)**.

- Humans appear and form themselves as **unique persons— not as mere objects** — through **speech and action in public space. (공적인 드러남 통해 고유한 인격적 존재됨)**

The Fundamental Condition of Publicness : The "Space of Appearance" 공공성의 근본 조건으로서 '드러남의 공간'

- Publicness is about forming and realizing the space where one **can appear before others**. (공공성은 타자에게 드러날 수 있는 가능성과 장 구성의 문제)
- It asks **who, how, and in what way individuals can appear in public**.

- **This Space is Continuously Reconstituted**
  - It is not a given place, but emerges through relational and performative acts. ('드러남의 공간'은 관계적 행위 속에서 성립)
  - Therefore, publicness must be renewed, criticized, and enacted continuously. (공공성은 지속적으로 갱신·비판·실천되어야)

Today: Technological Mediation of the "Space of Appearance"

- 오늘날은 드러남의 공간이 기술적 매개와 함께 구성

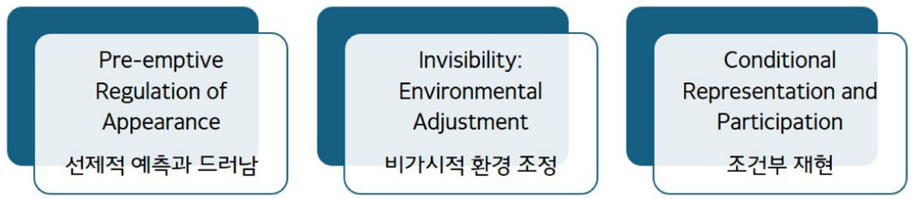
- **Algorithmic Mediation**
  - Automated algorithms preselect and adjust the conditions of appearance. (알고리즘이 드러남의 조건을 선별·조정)

- **New Question of Publicness**
  - **How is the "space of appearance" constructed and limited through technology? (**기술 매개 속에서 드러남의 공간은 어떻게 구성·제한되는가)

III. Automated Algorithms and the “Space of Appearance”

자동화된 알고리즘과 ‘드러남의 공간’

Tension Between Automated Algorithms and the "Space of Appearance
‘드러남의 공간’과 자동화된 알고리즘의 긴장 관계



Pre-emptive Regulation of Appearance
선제적 예측과 드러남

- Predicts outcomes and allocates resources to achieve or avoid them. (예측 결과를 바탕으로 자원 배분)
- Pre-selects who may appear by adjusting behavior before encounters occur.(조우 이전에 드러남의 가능성 선별·차단)
- Recommendation systems limit perception and interactions, shaping relational experience.(예: 추천 알고리즘의 인지·상호작용의 가능성 조정)
- Navigation algorithms may reinforce social stigma and spatial isolation. (공간적 낙인·고립 심화 가능성)

Invisibility: Environmental Adjustment
비가시적 환경 조정

- **Automated algorithms act not through visible control of individuals,** but by **adjusting environments that shape entire populations.** (개인을 통제하지 않고 인구집단의 환경을 조정)
- Example: In Songdo’s “Smart Village,” living spaces are intertwined with data collection and experimental infrastructures (MOLIT, 2020). (송도 스마트빌리지: 데이터·실험 인프라 결합된 환경)
- **These invisible adjustments integrate multiple layers — data, sensors, infrastructure —** forming **algorithmically governed environments.** (데이터·센서·인프라 결합하여 알고리즘 환경 형성)

Invisibility: Environmental Adjustment
비가시적 환경 조정

- In algorithmically governed spaces, **appearance itself becomes conditional.** (알고리즘적 공간에서 드러남은 조건적)
- **Those who perform algorithmic participation remain visible;** others fade into data absence.
(알고리즘이 요청하는 참여를 수행하는 존재만 가시화됨)
- Individuals are **not revealed as unique beings**, but **reproduced as data composites** that fit statistical models.
(고유한 인격이 아니라, 통계적 모델에 맞춰진 데이터 조합으로 재현됨)
- Visibility no longer expresses singular identity, but reflects algorithmic **representation and selection.**
(드러남은 고유한 자아 표현 아닌 알고리즘적 재현과 선별의 결과)

III. Future Task: Rethinking Publicness in the Age of Artificial Intelligence

인공지능 시대의 공공성, 향후 과제

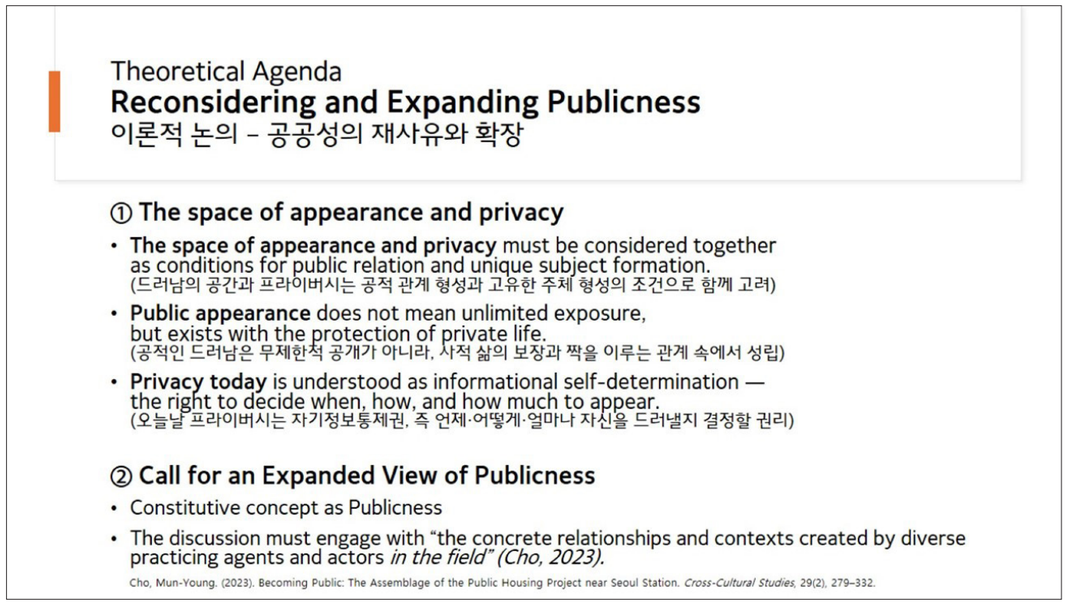
Opacity and the Question of Publicness in Algorithmic Space 불투명성과 알고리즘적 공간에서의 공공성 문제

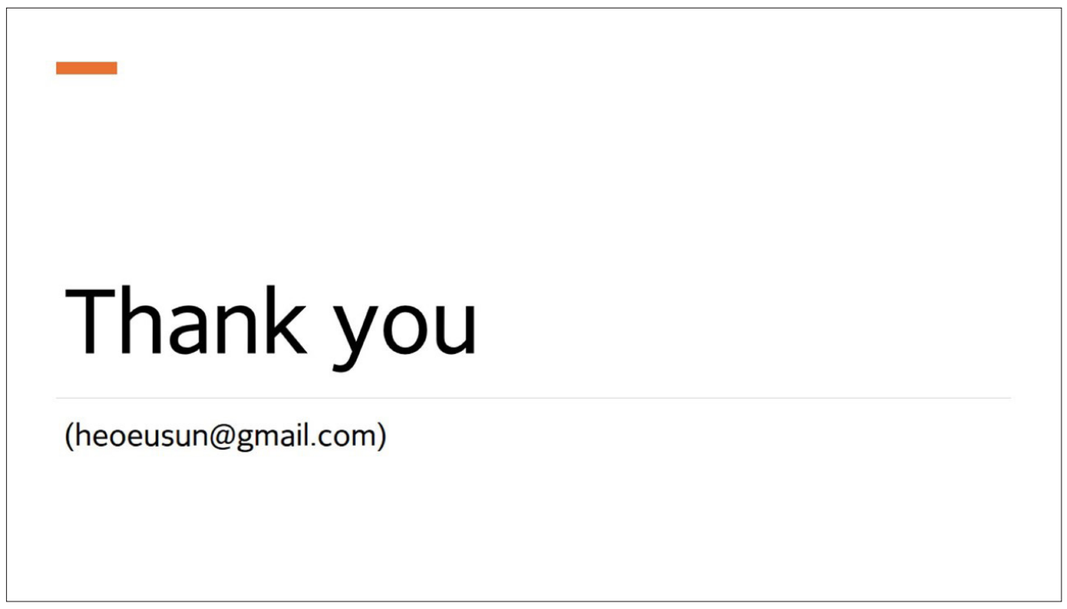
- **algorithmic opacity**, individuals find it difficult to recognize or critique the exclusionary structures of space.
(알고리즘 불투명성, 개인은 공간의 배제적 구조를 인식·비판하기 어렵다)
- The spaces mediated by automated algorithms are **not merely technical systems or applications**, but **problems of publicness** — determining who can appear and participate
(알고리즘이 매개하는 공간은 단순한 기술 적용의 문제가 아니라 '누가, 어떻게 드러날 수 있는가'의 공공성 문제)
- Thus, **critical reflection and practical response** to algorithmic operations lead to the central question:
How is publicness constituted and realized today?
(따라서 자동화된 알고리즘에 대한 비판적 성찰과 실천은 오늘날 공공성이 어떻게 구성·실현되는가의 물음으로 이어짐)

Practical Agenda
: Reconstructing Algorithmic Spaces as Public
실천 - 공적 공간으로서의 알고리즘 기반 공간 재구성

- **Reconstitute algorithmic environments as public spaces** where diverse individuals can appear and interact.
(알고리즘 기반 환경을 다양한 타자의 고유성이 드러나는 공적 공간으로 재구성)
  1. **Move beyond technological and economic efficiency →** toward publicness-centered design and governance.
(기술·경제적 효율성 중심 기획을 넘어 공공성 중심 설계와 거버넌스로 전환)
  2. **Expand citizen participation beyond data provision →** include policy-making, spatial design, and oversight processes.
(시민 참여는 데이터 제공을 넘어 정책·설계·감사 과정으로 확장되어야 함)

- **Remaining Tensions (공공성과의 긴장 관계, 지속적 성찰 필요)**
- Requires **continuous reflection and response** to the question: Who appears, and how?





인간과 AI의 공동창조: 지속 가능한 미래를 위한 자연지수(NQ)의 역할

Human-AI Co-creation: The Role of Nature Quotient in a Sustainable Future

호 만 퉁
베트남 사회과학원 철학연구소 연구원

Ho Manh Tung
Researcher, Institute of Philosophy, Vietnam Academy of Social Sciences

Abstract

This paper discusses the human-nature-technology nexus via two conceptual frameworks: the Nature Quotient and the seven premises of human-AI cocreation. Nature Quotient (NQ) is a proposed form of intelligence that measures a person's ability to understand, interact with, and live in harmony with the natural world. It goes beyond traditional measures like IQ (Intelligence Quotient) and EQ (Emotional Quotient) by focusing on ecological consciousness and sustainable behavior. The seven premises of human-AI cocreation present starting points for making sense the complex interplay of human-AI interactions with how we view ourselves, societies, and ecology. Integrating these two perspectives, we discuss how the ethics and design of AI systems can benefit from the rich yet fading reservoir of human natural wisdom (NQ) and moving us toward a more eco-surplus culture of human-AI interaction.

Introduction: Reconciling Human-AI Co-creation with Ecological Consciousness

Humanity's relationship with artificial intelligence (AI) is entering a new era of co-creation—one fraught with both transformative potential and ethical peril. As AI systems proliferate and entwine with daily life, scholars urge that guiding this human-AI partnership toward sustainability is fundamentally an ecological challenge. Nature Quotient (NQ) – defined as "a distinct, essential form of intelligence that enables humans to comprehend, adapt to and harmonize with complex natural systems" (Vuong & Nguyen, 2025) – has been proposed as a critical compass. By "fostering deeper ecological consciousness and guiding sustainable behavior," a high NQ "can counteract the anthropocentric biases inherent in conventional intelligence models and catalyze a sociocultural shift from an eco-deficit paradigm to an eco-surplus culture" (Vuong & Nguyen, 2025). In other words, cultivating NQ infuses human decisions with environmental attunement, counterbalancing AI's tendency to reflect human-centric, growth-driven values. Without such ecological consciousness, the co-evolution of humans and AI may reinforce harmful paradigms – a trajectory evidenced by rising "climate apathy" and denialism. Indeed, recent commentary warns that to overcome climate apathy, society must leverage advanced AI for good; "failing to do so risks leaving the power of these technologies to fall into the hands of climate change denialists" (Vuong & Ho, 2024). This essay advances the thesis that implementing the seven premises of human-AI co-creation is, at its core, an ecological quest – one requiring NQ as a moral and practical guide. In the following sections, we examine each premise in turn, arguing that NQ is indispensable to embedding multi-species, systemic, and long-term perspectives into our human-AI future. The alternative – proceeding without NQ – risks deepening ecological degradation and complacency even as technological intelligence soars.

The Primacy of Social Structures

Premise 1: The primacy of social structures. Any analysis of human-AI interaction must begin by recognizing that AI does not emerge in a vacuum; it is conditioned by pre-existing social and cultural contexts. As Ho and Vuong (2025) observe, "prior to any individual human and to any AI system, there have always been social, cultural, political, and historical structures that predispose individuals and machines to certain ways of thinking, behaving, and being" (Ho & Vuong, 2025). AI algorithms are trained on human-generated datasets and thus inevitably inherit the values and biases of their environments: "any algorithms that interact with humans are also trained with datasets that come from the preexisting social worlds" (Ho & Vuong, 2025). Crucially, many of these inherited structures reflect an anthropocentric worldview that elevates human priorities above ecological well-being. Contemporary analyses of AI ethics frameworks find a pervasive human-centered bias. For example, Rigley et al. (2023) report that "human-centred AI ethics standards tend to prioritise humans over nonhumans" and that this anthropocentrism implicitly "permit[s] harm to the environment and animals" and undermine[s] the stability of ecosystems (Rigley et al., 2023). Coeckelbergh (2025) similarly defines anthropocentric AI ethics as "concerned with how AI affects humans while ignoring the impact of AI on animals and the environment. It is human-centered." Such an approach "fails to sufficiently address planetary sustainability… the environment is seen as merely instrumental to human needs rather than a stakeholder with intrinsic value" (Coeckelbergh, 2025). In short, the social-structural status quo into which AI is born is one of human dominance, short-term economic growth imperatives, and disregard for non-human interests.

This recognition is vital because it reveals why Nature Quotient is so necessary at the foundation of human-AI co-creation. NQ represents the cognitive shift needed to challenge entrenched anthropocentrism. By definition, NQ "reflects people's capabilities to move beyond anthropocentric desires and shortsightedness to foster ways of holistic thinking, behaving and living that are harmonious with and nurturing toward nature" (Vuong & Nguyen, 2025). It functions as a corrective lens, helping both individuals and the AI systems they design to see beyond the immediate human-centric frame. Without NQ, AI will simply amplify existing biases encoded in data and institutions – a machine "socialized" into the same blind spots and exploitative habits that have driven ecological crises. The primacy of structures thus implies that if those structures are flawed (e.g. biased against the long-term, multi-species good), AI will magnify the flaw. Cultivating NQ at all levels – from AI developers and policymakers to algorithms themselves – is how we begin to rebuild those structures on ecologically sound principles. In sum, the first premise underscores that human-AI co-creation is not happening on a blank slate; it requires intentional reorientation of underlying social values. NQ provides that reorientation by embedding an ecological conscience into our collective intelligence from the ground up.

The Necessity of Co-creation

Premise 2: The necessity of co-creation. Human agency remains a central factor in the AI age: people will not passively accept a future dictated entirely by machines, no matter how "intelligent" those machines become. Co-creation – a collaborative partnership between humans and AI – is thus not just a nice ideal but a necessity arising from human nature itself. As Ho and Vuong (2025) explain, "while each human desires differing levels of agency and autonomy, all do have some desire for freedom. … [E]ven if we have the most moral and superhuman intelligent machines, it is hard to see the future where humans only obediently abide by what the machines instruct" (Ho & Vuong, 2025). In other words, attempts to impose top-down AI solutions without human participation or buy-in are likely to provoke resistance (conscious or unconscious). This "existential perversity" – the tendency for people to assert their freedom even against their own rational interests (Bloom, 2019, as discussed in Ho & Vuong, 2025) – means that sustainable outcomes cannot be achieved by AI alone or human

actors alone. They must be co-created through iterative human-AI interaction, with humans retaining a sense of ownership and moral agency.

Critically, true co-creation goes beyond humans merely using AI tools; it involves AI systems actively influencing human attitudes and vice versa in a continuous loop. Recent analyses suggest that human-AI interactions "extend beyond merely seeking and consuming information; they also involve a process of acculturation" (Vuong, 2023b, as cited in La et al., 2025). That is, AI is not a neutral instrument – it shapes our cognition and values as we shape its development. Ideally, this mutual shaping can be harnessed for ecological betterment. Vuong (2023b) envisions AI systems as "powerful allies in our efforts to prevent the catastrophic impacts of global warming", aiding humans in expanding their understanding of environmental issues and even "guiding us through content recommendations" that help "cultivate [a] new mindset (a set of core values) centered on ecosystem protection and restoration" (La et al., 2025, citing Vuong, 2023b). In practical terms, an AI chatbot might not only answer questions about climate change, but gently steer users toward pro-environmental knowledge and empathy, effectively co-creating a more sustainable culture. Such a symbiotic process exemplifies NQ in action: the AI embeds and disseminates ecological wisdom, while humans, in turn, train and refine the AI with feedback grounded in ethical and environmental priorities.

Co-creation is also needed because many problems at the human-AI nexus require hybrid intelligence – the combination of computational power with human judgment. Nguyen and Vuong (2025) argue that ensuring the quality of AI-generated knowledge will require "robust mechanisms to evaluate and verify AI-generated outputs," potentially including "hybrid human–AI oversight" so that what AI produces is not only accurate "but also contextually appropriate and ethically sound" (Nguyen & Vuong, 2025). In ecological governance, for example, AI might analyze vast climate data sets far better than any human, but decisions on adaptation strategies demand human ethical deliberation and local contextual knowledge. A collaborative approach leverages each party's strengths. Co-creation thus emerges as the only viable path to integrating AI into society's response to climate and sustainability challenges: humans and AI must learn from each other. The necessity of co-creation ultimately reflects a philosophical stance that technology should augment human wisdom, not replace it. By cultivating NQ, humans remain actively engaged as moral anchors (as we discuss next) and ensure AI's contributions are aligned with multi-generational, multi-species well-being. In short, co-creation guided by NQ transforms AI from a potential rival or opaque oracle into a cooperative partner in the project of sustaining life on Earth.

The Centrality of Context

Premise 3: The centrality of context. Human-AI co-creation does not occur in abstract generality; it is always grounded in specific contexts – cultural, historical, ecological – which critically shape meanings and outcomes. This premise builds on the idea that humans must remain the moral anchor of AI (addressed in the next section) by emphasizing the importance of situational awareness and sensitivity in all human-AI interactions. An "enlightened" or "awakened" human approach to AI, according to Ho and Vuong (2025), involves returning to and applying humanity's rich traditions of wisdom in each context: "Humans have developed a huge, rich repertoire of philosophy and science of human flourishing. The age of human–computer interactions will give us boundless possibilities, but to choose wisely among these, we must go back to study and apply these traditions and practices, both at personal and collective level" (Ho & Vuong, 2025). In essence, context includes our accumulated moral philosophies, cultural values, and place-based knowledge – resources that must inform how we direct AI. An AI recommendation that might be appropriate in one cultural or ecological context could be harmful or meaningless in another. Without human contextualization, AI's one-size-fits-all optimizations can misfire badly.

One planetary context cannot be ignored: the Earth itself. As theologian Hỏngkeun Kim (2025) argues, we must situate human–AI relations within the reality of a finite biosphere. Kim calls for a "post-anthropocentric understanding of cultural evolution from a planetary perspective", insisting that discussions of human–AI relationships "must transcend the narrow confines of human history" and "incorporate no less than the planetary system as the very field in which human–AI interactions unfold" (Kim, 2025). This perspective reminds us that every AI application (from data centers' energy usage to automated decisions affecting land use or wildlife) plays out within an ecological context. The centrality of context thus has a dual meaning here: AI must be contextualized not only in human social terms but also in ecological terms – the context of life. NQ contributes on both fronts. On one hand, a person with high NQ approaches problems with a systems view, aware that local actions have global ripple effects and that short-term gains must be weighed against long-term sustainability. On the other hand, NQ involves humility about one's own knowledge – recognizing the "contested, evolving, and frequently incomplete nature of knowledge" itself in any context (Nguyen & Vuong, 2025). This humility is crucial when facing AI outputs that may appear authoritative. Developing "validation and filtering mechanisms" for AI's information – akin to scientific peer review – is contextual work, discerning what fits the facts on the ground and ethical norms (Nguyen & Vuong, 2025).

Incorporating context also means respecting local and Indigenous knowledge and the values of communities who are directly impacted by environmental decisions. AI models trained on global data might overlook context-specific variables or cultural values. For example, an AI-driven conservation plan must consider indigenous peoples' relationships to the land, not just biodiversity metrics. High-NQ individuals and societies will insist that AI's role remains that

of an advisor whose suggestions are interpreted through human contextual understanding, rather than an oracle dictating context-insensitive edicts. By making context central, we ensure AI augments situated human judgment rather than overriding it. This aligns with the ethical principle of justice as recognition: including nature and non-human actors as stakeholders. Indeed, "expanding the scope of AI ethics beyond an anthropocentric focus" to include "more-than-human actors such as animals and ecosystems" is now seen as imperative (van Uffelen et al., 2025). In summary, the third premise highlights that where and how we deploy AI is as important as what the AI does. Through NQ, humans maintain a contextual lens, integrating traditional wisdom and environmental realities into the co-creative process with AI.

The Human as a Moral Anchor

Premise 4: The human as a moral anchor. No matter how advanced AI becomes, humans must remain the ultimate source and arbiter of moral values in the human-AI relationship. Technology can aid ethical decision-making, but it cannot supplant the deep moral reasoning and empathy that arise from human consciousness and cultural evolution. Ho and Vuong (2025) emphasize that the foundational ideals – Truth, Goodness, Beauty – in human-computer interactions depend on an "awakening" of human judgment and virtue. They ask, "What does it mean to be enlightened or awakened as a human being?", noting that this question has been the pursuit of philosophers and spiritual leaders for millennia. In the AI era, they argue, we have to actively draw on humanity's moral heritage: "The age of human–computer interactions will give us boundless possibilities, but to choose wisely among these, we must go back to study and apply these traditions and practices" at both the individual and societal level (Ho & Vuong, 2025). In practical terms, this means that human values, refined through centuries of ethical thought and lived experience, should guide the development and deployment of AI. Rather than expecting AI to be moral on its own, humans must teach and frame AI with moral considerations – effectively anchoring it to a moral compass.

Nature Quotient offers a crucial moral dimension often missing from conventional IQ or EQ-driven frameworks: an ecological ethic. Vuong and Nguyen (2025) describe NQ as not only an intellectual skill set but "a corrective force against the potential hubris" associated with high cognitive or emotional intelligence unchecked by humility. They caution that while IQ and EQ can foster innovation and empathy, they may also "contribute to a sense of human superiority or detachment from the natural world" (Vuong & Nguyen, 2025). NQ, by contrast, "embeds intelligence within a framework of humility, interdependence and systems thinking". It "reorients personal and collective intelligence toward more grounded, ethical and sustainable behaviors by anchoring knowledge and emotional insight in a deep awareness of ecological interconnection" (Vuong & Nguyen, 2025, emphasis added). In essence, NQ grounds our moral reasoning in the reality that humans are part of, not apart from, the community of life. This broadened moral circle – extending care and intrinsic value to non-human beings and future generations – is exactly what is needed to anchor AI's trajectory in sustainability.

Consider a concrete example: an AI system managing a supply chain might efficiently cut costs (an IQ objective) and even optimize worker satisfaction (an EQ-related goal), yet still source materials in a way that devastates rainforests or exploits underpaid miners. A human leader with high NQ acting as a moral anchor would insist the AI also account for environmental and social justice factors, reflecting an awakened conscience that sees beyond immediate profits. Without such human intervention, the AI, per its training, might unconsciously perpetuate a narrow value function (e.g. short-term efficiency) divorced from ethical context. History offers sobering lessons of advanced knowledge deployed amorally – from atomic science to Big Data. The presence of an "awakened" human in the loop – one who prioritizes Truth (honest data), Goodness (ethical outcomes), and Beauty (harmony with nature) – is what can prevent AI from becoming merely a force-multiplier of greed or negligence. In moral philosophy terms, humans must impose deontological and virtue-ethics considerations on AI's largely utilitarian logic. By cultivating NQ across society, we effectively raise a generation of decision-makers who serve as moral anchors, ensuring human-AI co-creation serves life-affirming ends rather than undermining the very ecological foundations of morality itself.

The Emergence of New Values

Premise 5: The emergence of new values. Human-AI co-creation is a dynamic process that will generate new values, norms, and cultural patterns – some of which may be unforeseen or unintended. The integration of AI into daily life can amplify certain ideals while attenuating others, leading to what Ho and Vuong (2025) call "acculturative emergence." They note that "from human–computer interactions, there emerge values that, whether being consciously understood and examined or not, will be incorporated into the core values of human society." In many cases, values that were once marginal or dormant can resurface at scale through billions of algorithm-mediated interactions: "values arising from the billions of interactions between humans and algorithms in the hyperweb, once marginal or forgotten, can become mainstream again" (Ho & Vuong, 2025). For example, the widespread use of social media (an earlier AI-driven transformation) unexpectedly brought values like communal sharing of information to the fore, but also negative norms like performative outrage. Likewise, as generative AI becomes ubiquitous, it might normalize new attitudes toward creativity, privacy, or truth. The key point is that human-AI co-creation doesn't just implement existing values – it actively shapes and creates values. This realization carries both promise and peril.

On the one hand, we have the opportunity to deliberately cultivate and mainstream ecocentric values through human-AI interactions. If guided by NQ, AI could help revive and spread cultural ideals that respect nature. For instance, Auerbach (2023) observed that the emergent dynamics of online networks can undermine classical values like democracy or inclusivity despite the personal intentions of users or tech owners. By the same token, deliberate design of human-AI systems could elevate long-submerged ecological values to prominence. Vuong and Nguyen (2023) offer a vivid example in their Kingfisher essay: symbolic representations of nature can reconnect people with forgotten values of coexistence. The authors argue that images and stories of the kingfisher bird can "gradually build humans' perceived values of the natural world," helping to transform "human value systems from eco-deficit to eco-surplus mindsets" (Vuong & Nguyen, 2023). Here, wildlife and art serve as mediators of value emergence, but AI could amplify this process – for example, by curating nature-inspired content, facilitating citizen science that highlights biodiversity, or personalizing education about ecological interdependence. Indeed, exposure to nature (direct or AI-mediated) often triggers what Vuong et al. (2025) term "serendipitous moments where people come to realize previously unexpected values of nature," leading to "the formation and internalization of knowledge and values that support environmental protection and restoration" (Vuong et al., 2025). An AI that increases our exposure to such moments – for instance, by simulating the consequences of climate inaction or by revealing hidden natural beauty – can sow the seeds of new, greener values in society's consciousness.

On the other hand, without the guiding hand of NQ, the new values emerging from human-AI entanglement could just as easily be problematic – reinforcing consumerism, complacency, or misinformation. As noted earlier, algorithm-driven interactions have already normalized short attention spans and echo chambers of thought. In the environmental domain, an unguided AI might inadvertently foster "climate apathy" by inundating users with disheartening news or by optimizing for engagement in ways that favor trivial content over sustainability insights. We have to proactively shape the character of value emergence. High-NQ individuals and communities would seek to design AI platforms that celebrate sustainable living, highlight long-term consequences, and make systemic connections salient. For example, a social media AI could be tuned (with human oversight) to boost content about successful regenerative projects or to connect people emotionally to endangered species, rather than solely pushing material that triggers short-term emotion or consumption. The emergence of new values in the age of AI is inevitable – what those values are is up to us. By infusing the co-creative process with NQ, we tilt the balance toward values of conservation, interdependence, and responsibility, rather than exacerbating the existing eco-deficit values of extraction and instant gratification. As Vuong and Nguyen (2025) succinctly frame it, "environmental protection and regeneration are not peripheral moral choices but existential necessities" – a truth that must percolate into the emergent value system of our AI-mediated civilization. Human-AI co-creation, approached wisely, can be a vehicle for that necessary cultural evolution.

The Inevitability of Change

Premise 6: The inevitability of change. The interactions among humans, AI, and the environment form a complex triangle, and change in one corner inevitably induces change in the others. As Ho and Vuong (2025) put it, "human–computer interactions are already changing and will continue to change human-to-human and human–nature interactions." These three spheres – HCI (human-computer), HHI (human-human), and HNI (human-nature) – are deeply intertwined, and striking a healthy balance among them is crucial. Ho and Vuong argue that "finding the balance among these three overlapping spheres of interactions would allow for human flourishing," and thus researchers and policymakers must "account for the impacts [AI] has on human–human and human–nature interactions" when assessing AI's effects (Ho & Vuong, 2025). In other words, as we integrate AI deeper into society, we must anticipate and manage cascading changes in social relations and in our relationship with nature. Change is not optional; it is already underway. The choices before us are about direction and magnitude, not about standing still.

One major change in recent decades is the rapid digitalization of human experience, which has profound implications for how we perceive and value nature. Vuong et al. (2025) observe that "the world is becoming increasingly urbanized and digitalized, leading to a growing disconnection between humans … and nature." Importantly, "this disconnection is perceptual rather than physical, as natural ecosystems provide the indispensable resources and services essential to human existence" (Vuong et al., 2025). People – especially younger, "digital-native" generations – may feel estranged from nature simply because their attention is captured by screens and virtual environments. AI, in its current trajectory, could worsen this trend: "efforts to isolate the digital world – particularly AI – from the natural world may inadvertently deepen the perceptual disconnection between humans and nature" (Vuong et al., 2025). However, AI can also be part of the solution. The same authors suggest that "AI should be recognized and utilized as a tool to disseminate information, raise awareness, provide education, and inspire individuals – especially urban and digital-native populations – to reconnect with and explore nature" (Vuong et al., 2025). This points to a proactive strategy: use AI's vast informational reach and personalization capabilities to bridge the human-nature gap. For instance, augmented reality apps could encourage city dwellers to engage with local biodiversity, or AI tutors could weave ecological literacy into everyday learning. By doing so, we leverage the inevitability of technological change to catalyze positive environmental change in attitudes and behaviors.

Change is also inevitable in the environmental sphere itself – climate systems are already

destabilizing, biodiversity is declining – and these changes will, in turn, force transformations in human society and technology. Lenton et al. (2019) warned of approaching climate tipping points, and indeed 2024 saw the global average temperature exceed the 1.5°C threshold for the first time in a single year (La et al., 2025). In facing such upheavals, complacency is not an option. Human-AI co-creation must adapt continuously; our ethical frameworks, policies, and technologies will all need updating as new realities emerge. Crucially, NQ can help humans and AI alike navigate uncertainty by emphasizing adaptability and resilience. Granular Interaction Thinking Theory (GITT), which underpins NQ, holds that survival "depends on the efficient management of information: acquiring it, storing it, transmitting it and processing it" in the face of environmental change (Vuong & Nguyen, 2025). In practice, this means fostering AI that is flexible, context-aware, and responsive to real-world feedback – not rigidly programmed for a world that no longer exists. It also means cultivating human mindsets that embrace change as an opportunity for growth rather than a threat to be ignored. We have seen that change is a double-edged sword: it can lead to loss (e.g. loss of traditional connections to nature) or to renewal (e.g. emergent eco-surplus values). The outcome will depend on how consciously and wisely we guide the co-evolution of human, AI, and ecological systems. A Nature-Quotient-informed approach treats change as inevitable but malleable: we cannot stop the tide, but we can set our sails to harness the winds of change toward a just and sustainable future.

The Importance of Proactivity

Premise 7: The importance of proactivity. Given the rapid pace of AI advancement and the urgency of ecological crises, waiting for problems to resolve themselves is a folly we cannot afford. We must actively chart a new path forward, anticipating challenges and acting early to steer human-AI co-creation in a life-affirming direction. As one parable from The Kingfisher Story Collection cautions, "There must be a plan of action because delaying will be dangerous." (Vuong, 2022). This ethos of foresight and timely action is echoed by scientists and ethicists alike. Jared Diamond (2011) warned that societies often collapse not from sudden external blows, but from a slow, self-inflicted erosion of their environmental support systems – essentially, a failure to act in time. In the context of AI, Vuong et al. (2025) similarly argue that "given the tremendous power of AI, especially when combined with humans' serendipity capability, its development and use must be guided by the utmost prudence and ethical considerations. Otherwise, the misuse of the immense power of AI … could steer humanity toward its destruction." They specifically note that humans, absent external threats, "may inadvertently pave the way to self-destruction by degrading the ecological systems that sustain them" (Vuong et al., 2025). These stark warnings convey a simple truth: the default trajectory of human-AI-techno-economic systems is not necessarily benign. Without deliberate intervention, AI will be driven by short-term profit and convenience, and the environment will be an afterthought – until crises force painful corrections. Proactivity means taking control of the narrative now, embedding ethical guardrails and sustainability objectives into AI by design rather than as an after-the-fact patch.

What does proactivity look like in practical terms? It includes policy and governance innovations, such as the "semiconducting principle" Vuong et al. (2025) advocate – a rule that environmental value can convert into economic value, but not vice versa, to prevent the market from commodifying nature without limits. It includes reorienting education to raise NQ alongside IQ, so that upcoming generations naturally integrate ecological thinking in their innovations. It includes interdisciplinary research agendas that forecast AI's long-term impacts on climate, biodiversity, and social equity, rather than narrowly focusing on near-term performance metrics. In the AI development community, proactivity might involve setting industry standards for energy efficiency and carbon footprint of AI models (Brevini, 2020; van Uffelen et al., 2025) and incorporating "planetary boundaries" into AI's optimization criteria. It certainly involves inclusivity and dialogue – ensuring that voices from the Global South, Indigenous communities, and future generations (voiced through proxy representations) are at the table when designing our AI-enhanced future. As Coeckelbergh (2025) notes, a truly global and morally adequate AI ethics would be "inclusive, open, respectful, dialogical, and truly relational," sensitive to differences and focused on the welfare of the planet rather than just the prerogatives of the powerful.

Proactivity also demands a certain humility and flexibility. We must be ready to course-correct as new information emerges – a principle well-understood in adaptive environmental management. This resonates with the scientific principle of intellectual humility highlighted by Rovelli (2018): not to blindly trust in past knowledge or models, but to constantly test and update them. In AI terms, it means monitoring outcomes and being willing to pull back or modify technologies that show unintended harm. It also means investing in "early warning systems" – whether computational (AI predicting tipping points) or social (watchdog groups tracking AI's societal impacts) – so that we are not blindsided. The overarching point is that time is of the essence. Environmental change is accelerating, and AI development is on exponential curves; our ethical and governance response must not be linear and lethargic. Cultivating NQ in tandem with AI gives us the best chance to stay ahead of the curve, because high-NQ actors will inherently be scanning the horizon for ecological ramifications and moral dilemmas. They will, one hopes, have the wisdom to heed the proverbial canary in the coal mine – or perhaps, to use a more apropos metaphor for our context, the kingfisher by the riverside alerting us to the health of our waters. In proactive co-creation, we listen to those alerts and act decisively, rather than hitting the metaphorical snooze button on warnings of climate and AI risks.

Conclusion: Charting a New Path Forward

The co-creative journey of humans and AI is ultimately a story yet to be written – a story whose ending will be determined by the values we adopt and the actions we take today. This essay has argued that implementing the seven premises of human-AI co-creation is fundamentally an ecological challenge, demanding the guidance of Nature Quotient (NQ) at every step. We began by recognizing the primacy of social structures: AI will mirror the anthropocentric, exploitative biases of its upbringing unless we consciously infuse new ecological values. We identified the necessity of co-creation, seeing that human autonomy and AI capability must partner, not clash – and that NQ can harmonize this partnership by aligning it with the natural world's needs. We highlighted the centrality of context, urging a planetary perspective and local grounding to ensure AI serves the whole earth community, not just abstract metrics. We affirmed the role of the human as a moral anchor, asserting that our time-tested wisdom traditions and an awakened ecological conscience must guide ultra-rational AI toward ethical ends. We examined the emergence of new values, hopeful that with NQ these might bend toward an eco-surplus culture rather than deeper eco-deficit. We acknowledged the inevitability of change – social and environmental – and the need to use AI as a tool to adapt and even reconnect, rather than further disconnect, humans from nature. Finally, we underscored the importance of proactivity: the urgency of acting now, deliberately and prudently, to set a sustainable course for human-AI co-evolution.

Underpinning all these arguments is a unifying insight: Our relationship with AI cannot be separated from our relationship with the rest of nature. If we treat AI as just another instrument for short-term human gain, we will simply accelerate environmental collapse – automating ourselves into oblivion. But if we cultivate a high collective NQ, we position AI as a powerful ally in healing the planet and securing a flourishing future for all species. NQ enables AI to embed multi-species, systemic, and long-term perspectives, helping us transcend the myopia that has plagued the industrial-growth era. It reminds us, as Deep Ecology does, that every form of life has intrinsic value and that humanity is but one strand in the web. In practical terms, an NQ-guided AI would measure success not by GDP or click-through rates alone but by metrics of ecosystem health, carbon reduction, and community well-being. It would highlight interconnections – how consumer choices affect distant forests, how climate impacts drive human migration – thereby educating and nudging us toward wiser choices. It would also foster empathy beyond the human realm, perhaps through simulations of animal perspectives or by illuminating the beauty of natural processes, countering what some have called the "extinction of experience" in urban populations (Soga & Gaston, 2016).

Conversely, proceeding without NQ – leaving AI development solely in the hands of market forces or narrow definitions of efficiency – risks entrenching what some researchers term "eco-deficit culture" and even a sense of fatalism or apathy about our climate destiny. Vuong and Ho (2024) implore us to "escape climate apathy by harnessing the power of generative AI," suggesting that the same tools causing information overload can be repurposed to inspire action. This captures a broader principle: technology is what we make of it. The role of NQ is to ensure we make the human-AI symbiosis a force for regeneration, not destruction. It acts as an ethical North Star, continually pointing us back to the understanding that our fate is intertwined with the earth's fate. As Kim (2025) observes, the traditional hierarchy that placed humans above all is no longer tenable; modern science and ecology reveal "entanglement, symbiosis, interdependence" at every level of life, such that "human and nonhuman creatures… each respectively contribute to the co-creation of an ever-new Earth" (Kim, 2025). In this light, human-AI co-creation is simply the latest chapter in the story of life co-creating with life – except now one partner is an artifact of human ingenuity. Whether this artifact undermines or uplifts the broader community of life will depend on the wisdom we imbue in it.

Charting a new path forward thus means making a civilizational choice: do we continue on the path of "claiming nature as our own", or do we finally recognize nature as a co-equal partner (Vuong, 2023a)? The former path leads to what some have called the Anthropocene's dead end – a world of machines and humans clinging to dwindling resources. The latter path imagines an Ecocene or Symbiocene (Rose, 2013) where advanced technology operates within ecological guardrails and where development is measured by healing and balance, not extraction. The seven premises outlined – social structures, co-creation, context, moral anchoring, value emergence, change, and proactivity – offer a roadmap to reach that Symbiocene future. Each premise is a reminder that our tools and systems must ultimately serve life. By implementing them with a high Nature Quotient, we ensure that our intelligence – whether biological or artificial – always loops back to its ethical foundation: the thriving of this wondrous, interdependent web that we call home.

References

Adamson, J. (2013). Nature's Call: Reconnecting Humanity with the Natural World. [Definition of Nature Quotient informed by Deep Ecology as exposure to raw nature].

Airoldi, M. (2021). Machine Habitus: Toward a Sociology of Algorithms. Wiley. [Explores how algorithms are "socialized" by human data feedback, reproducing social structures].

Bloom, P. (2019). The strange appeal of perverse actions. The New Yorker. [Concept of "existential perversity" in human behavior].

Brevini, B. (2020). Is AI Good for the Planet? (Chapter in Amazon: Understanding a Global Communication Giant). [Discusses the often-overlooked environmental footprint of AI and calls for sustainable AI practices].

Coeckelbergh, M. (2025). Three challenges for a global AI ethics: towards a more relational normative vision. AI and Ethics. 5(1), 13–30. [Argues that anthropocentrism and lack of relational thinking hinder AI ethics, proposing a more inclusive, planet-centered approach].

Diamond, J. (2011). Collapse: How Societies Choose to Fail or Succeed. Penguin. [Warns that societies can self-destruct by depleting environmental resources].

Ho, M.-T., & Vuong, Q.-H. (2025). Five premises to understand human–computer interactions as AI is changing the world. AI & Society, 40(1), 1161–1162. [Outlines five key premises (primacy of structures, human autonomy, awakened humans, emergent values, and interactional balance) for analyzing HCI in the age of AI].

Kim, H. (2025). Artificial intelligence and the sustainable future of co-creation. Zygon: Journal of Religion and Science, 60(2), 389–408. [Critiques human-centered AI discourse and advocates a post-anthropocentric, planetary perspective on human–AI relations, emphasizing co-creation and interdependence].

La, V.-P., Nguyen, M.-H., Tran, T. T., & Vuong, Q.-H. (2025). Are we on the right track for mitigating climate change? Visions for Sustainability, 24(Special Issue), 1–51. [Reviews global climate efforts and argues for shifting from an eco-deficit to an eco-surplus culture, highlighting the need for value and cultural transformation].

Nguyen, M.-H., & Vuong, Q.-H. (2025). Navigating the new landscape of knowledge in the age of generative AI. AI & Society, 40(2), 555–558. [Discusses challenges of generative AI for knowledge integrity and proposes human–AI oversight to ensure outputs are contextually appropriate and ethically sound].

Rigley, E., Chapman, A., Evers, C., & McNeill, W. (2023). Anthropocentrism and environmental wellbeing in AI ethics standards: A scoping review and discussion. AI, 4(4), 844–874. [Finds that most AI ethics guidelines are human-centered (anthropocentric) and discusses how this focus permits harm to nonhumans and ecosystems].

Rupprecht, C. D. D., et al. (2020). Multispecies sustainability. Global Sustainability, 3, e34. [Introduces the concept of "multispecies sustainability," arguing that true sustainability requires meeting the interdependent needs of all species, present and future].

Vuong, Q.-H. (2022). The Kingfisher Story Collection. AISDL. [A collection of fables and stories (e.g., "Bogeyman," "GHG Emissions") used in scholarly contexts to illustrate principles of wisdom, scientific ethics, and environmental conscience].

Vuong, Q.-H. (2023a). Meandering Sobriety. AISDL. [Essays prompting reflection on human folly and humility; quoted for the question "Will we ever stop claiming nature as our own?"].

Vuong, Q.-H. (2023b). Mindsponge Theory. De Gruyter. [Explores a theoretical framework of information processing and value formation in the human mind; referenced regarding acculturation via information absorption].

Vuong, Q.-H., & Ho, M.-T. (2024). Escape climate apathy by harnessing the power of generative AI. AI & Society, 39(6), 3057–3058. [Short commentary urging proactive use of AI to engage the public on climate change and overcome apathy, rather than allowing AI to be used by climate denialists].

Vuong, Q.-H., & La, V.-P. (2025). On serendipity and non-linear thinking in addressing environmental conundrums. Sustainability: Science, Practice and Policy, 21(1), 2480428. [Emphasizes the role of serendipitous discovery and open-minded innovation (aided by AI) in solving climate and environmental challenges, while warning of ethical pitfalls].

Vuong, Q.-H., & Nguyen, M.-H. (2023). Kingfisher: Contemplating the connection between nature and humans through science, art, literature, and lived experiences. Pacific Conservation Biology, 30(4), 361–366. [Uses the kingfisher bird as a symbol to discuss human-nature disconnection and argues for cultural pathways (science, art, indigenous knowledge) to rebuild an "eco-surplus" culture].

Vuong, Q.-H., & Nguyen, M.-H. (2025). On Nature Quotient. Pacific Conservation Biology, 31(1), 1–7. [Defines and elaborates the concept of Nature Quotient (NQ) as the capacity to understand and harmonize with complex natural systems; discusses how NQ counters anthropocentrism and can drive a shift to an eco-surplus culture].

van Uffelen, N., Lauwaert, L., Coeckelbergh, M., & Kudina, O. (2025). Towards an environmental ethics of artificial intelligence. Journal of Responsible Technology, 11, 100050. [Examines AI's environmental impacts and argues that expanding AI ethics to consider ecosystems and non-human entities moves us beyond anthropocentrism, proposing criteria for environmentally just AI design].

분과회의 세션 02
Parallel Session

AI와 인문학연구
AI and Humanities Research

AI 시대 인문학 연구 거버넌스

Humanities Research Governance in the AI Era

김바로
한국학중앙연구원 교수

Baro Kim
Professor, The academy of Korean Studies

초록

본고는 생성형 AI 시대, 인문학 연구의 위기를 진단하고 그 대안으로 인문 데이터 중심 연구 거버넌스를 제안한다. 이를 위해 KCI 데이터 분석과 공공 DB 사례 분석를 수행했다. 1) KCI 분석 결과, 국내 AI 담론이 데이터 편찬이라는 핵심 논의 없이 양적으로만 팽창했음을 밝혔으며, 2) 공공 DB 사례 분석을 통해 이 학문적 공백이 현실에서는 데이터가 고립된 데이터 갈라파고스(Data Galapagos)를 형성하는 구조적 실패로 귀결되었음을 실증했다. 결론적으로 인문학이 AI 시대의 주체로 서기 위한 3대 원칙을 제안한다. 첫째, 기술·의미론적 실패를 극복하기 위한 인문학 주도의 데이터 편찬과 정당한 보상, 둘째, 신뢰할 수 있는 AI를 위한 다원적 관점의 시맨틱 데이터 구축, 셋째, 제도적·정책적 실패를 넘어선 지속가능한 거버넌스 구축과 글로벌 연대이다.

Abstract

This paper diagnoses the crisis facing humanities research in the era of generative AI and proposes a humanities data-centric research governance as a solution. To this end, it conducted KCI data analysis and public database case studies. 1) The KCI analysis revealed that domestic AI discourse has expanded only quantitatively, without a core discussion on data compilation. 2) The public DB case studies demonstrated that this academic vacuum has resulted in a structural failure, creating isolated Data Galapagos.
Based on this diagnosis, this paper proposes three principles for the humanities to establish itself as a primary agent in the AI era: first, humanist-led data compilation and just reward to overcome tech-semantic failures; second, the construction of pluralistic semantic data for trustworthy AI; and third, the establishment of sustainable governance and global solidarity to move beyond institutional and policy failures.

1. 서론[1)]

최근 생성형 인공지능(AI)에게 "한국 전통 회화 양식으로 이순신 장군을 그려달라"고 요청하면, 종종 일본풍의 갑옷을 입거나 국적 불명의 인물이 등장하는 결과물을 마주하게 된다. 한때 유행했던 "세종대왕 맥북 프로 던짐 사건"[2)]과 같은 명백한 환각(Hallucination) 현상은 기술의 발전으로 점차 줄어들고 있지만, 이순신 장군의 사례처럼 특정 문화적 맥락과 전문적 지식이 요구되는 영역에서는 여전히 심각한 왜곡이 발생하고 있다. 이는 단순히 학습 데이터의 양이 부족하기 때문만은 아니다. 설령 개별 데이터가 존재할지라도, 각 정보의 역사적, 시각적, 문화적 맥락이 체계적으로 연결된 지식 구조, 즉 거버넌스(Governance)가 부재하기 때문이다.

AI 기술은 이제 인문학 연구의 도구를 넘어, 연구의 존재 이유까지 묻는 강력한 행위자(actor)로 부상했다. 특히 생성형 AI의 시대가 본격화되면서, AI가 생산하는 정보의 신뢰성 문제는 학계의 가장 시급한 과제로 떠올랐다. AI는 방대하고 정제되지 않은 웹 데이터를 기반으로 학습하기에, 그럴듯해 보이지만 검증되지 않은 결과물을 양산할 위험을 내포한다. 이러한 상황은 역설적으로 신뢰할 수 있는 원천 데이터(RAWDATA) 혹은 전문가에 의해 체계적으로 편찬된 지식그래프의 가치를 그 어느 때보다 중요하게 만들고 있다. AI에게 질문하는 법과 그 답변의 근거가 될 지식 체계를 설계하는 것, 이것이 바로 인문학에 주어진 새로운 시대적 과제이다.

이러한 거대한 전환 속에서 인문학계는 AI를 어떻게 활용할 것인가라는 도구적 논의를 넘어, AI 시대에 지속가능한 연구 생태계를 구축하기 위해 어떤 규칙과 협력 체계를 만들어야 하는가라는 근본적인 질문에 답해야 한다. 여기서 거버넌스란 단순히 데이터를 통제하고 관리하는 기술적 절차를 의미하지 않는다. 그것은 데이터의 생산, 유통, 활용, 보상에 이르는 전 과정에 걸쳐 "누가, 어떤 원칙으로, 어떻게 참여하고 결정할 것인가"를 다루는 공동체의 규범과 철학을 포함하는 포괄적인 사회-기술적 시스템이다.

본고는 이러한 문제의식 해결을 위해, 첫째, 한국학술지인용색인(KCI) 데이터를 활용하여 지난 20년간 국내 AI 인문학 담론의 지형도를 계량적으로 분석한다. 둘째, KCI 분석 결과가 현실의 데이터 인프라에 어떻게 투영되었는지를 규명하기 위해, 국내 주요 공공 데이터베이스의 사례를 분석한다. 마지막으로, 이 두 가지 진단을 종합하여, 파편화된 데이터를 연결하고 인문학의 사회적 역할을 재정립하기 위한 미래지향적 인문 데이터 중심 연구 거버넌스를 제안하는 것을 목표로 한다.

2. KCI 데이터로 본 AI 인문학 연구 지형(2004-2024)

본고가 제기하는 데이터 중심 거버넌스의 필요성은 추상적 담론이 아닌, 대한민국 학술 지형에 대한 구체적인 데이터 분석에 기반한다. 연구 데이터는 KCI 등재 논문 88만여 건 중 AI 관련 핵심 키워드[3)]를 포함하는 10,698건의 연구를 1차 대상으로, 이 중 데이터 관련 키워드[4)]를 포함하는 4,649건의 연구를 2차 심층 분

1) 본 발표문은 연구자의 프롬프트를 기반으로 Gemini 2.5 Pro(2025.10.10.)가 생성한 초고를 연구자가 직접 수정·보완하는 방식으로 작성되었다. 본 연구에서 제미나이는 초기 자료 수집과 관련 연구 동향 요약, 초고의 논리 구조 제안 등에 활용되었으며, 연구자는 제기된 핵심 아이디어를 비판적으로 검토하고, 전체 논지를 재구성하며 최종 집필을 책임지는 방식으로 협업을 진행했다.

2) 조선일보. "세종대왕의 맥북 던짐 사건에 대해 알려줘" 했더니 챗GPT가 내놓은 답변은?. 2023.03.05. https://www.chosun.com/national/weekend/2023/03/04/HR457QM36JFTXDUVAMMNG23MHQ/

3) 인공지능, Artificial Intelligence, AI, 기계학습, Machine Learning, 딥러닝, Deep Learning,자연어처리', NLP, Natural Language Processing, 생성형 AI, Generative AI, 챗GPT, ChatGPT, 거대언어모델, LLM, Large Language Model

4) 데이터, data, 거버넌스, governance, 기계가독, machine readable, machine-readable, 아카이브, archive

석 대상으로 삼았다.

KCI 데이터를 분석한 결과, 국내 AI 관련 연구는 외부 기술 충격에 민감하게 반응하며 양적으로는 크게 팽창했으나, 질적으로는 특정 응용 분야에 편중되고 핵심적인 논의가 부재한 불균형적 성장의 특징을 명확하게 보여주었다.

2.1. 거시 동향

국내 인문사회 분야의 AI 연구는 두 번의 뚜렷한 변곡점을 거치며 폭발적으로 증가했다. 2016년 알파고 쇼크 이전까지 연간 100건 미만에 머물던 논문 수는 2017년 400건을 돌파했으며, 특히 ChatGPT가 대중화된 2023년 이후 그 증가세는 더욱 가팔라져 2024년에는 연간 2,700건을 넘어서는, 전체 논문 중 5.8%가 AI를 언급하는 압도적인 양적 팽창을 보였다. 이는 학계의 연구 의제가 내재적 필요나 학문적 성찰에 의해 선제적으로 형성되기보다는, 외부의 거대 기술 이벤트가 발생한 후에야 비로소 담론이 활성화되는 기술 추종적 경향이 있음을 명확히 보여주는 지표이다.

이러한 양적 성장 이면에는 더욱 심각한 구조적 편중이 존재한다. 연구 분야를 세부적으로 살펴보면, 논의를 주도하는 학문 분야는 법학(1,048건), 교육학(913건), 경영학(778건) 순으로 나타났다. 이는 AI 기술의 사회적 도입에 따른 법적·윤리적 쟁점, 교육 현장에서의 활용, 산업 환경의 변화라는 당면 과제에 학계가 민감하게 반응하고 있음을 보여준다. 하지만 동시에 이는 역사학, 철학, 문학 등 인문학 고유 분과들이 상대적으로 논의의 중심에서 비켜나 있음을 방증한다. 즉, AI를 "어떻게 활용할 것인가?"라는 응용적·도구적 질문이 "AI 시대에 인간과 세계를 어떻게 이해할 것인가?"라는 인문학의 근본적인 성찰적 질문을 압도하고 있는 것이다.

2.2. 담론의 공백

데이터 거버넌스 관련 논의는 전체 AI 연구의 약 43.5%(4,649건)를 차지하며, AI 담론의 가장 중요한 하위 그룹을 형성하고 있었다. 그러나 그 내용을 심층적으로 들여다보면, 현재의 논의가 가진 결정적인 공백이 드러난다.

인공지능 전체 담론은 "ChatGPT", "4차 산업혁명" 등 화려한 기술의 결과와 사회적 영향에 대한 거시적 논의에 집중되어 있다. 그러나 정작 AI의 작동을 가능하게 하는 가장 근본적인 전제, 즉 데이터를 "누가, 어떻게 편찬하고 설계할 것인가?" 라는 인문학의 고전적이면서도 핵심적인 질문은 상대적으로 소외되어 있다. 데이터 거버넌스 담론에서 "개인정보", "블록체인"과 같은 키워드가 부상하는 것은 데이터의 관리와 통제에 대한 관심이 존재함을 보여주지만, 이는 AI가 만들어낸 결과물에 대한 사후적 대응(법적 규제, 신뢰성 확보)에 가깝다.

AI의 지적 수준과 방향성을 결정하는 데이터 편찬이라는 선행적이고 창조적인 과정 자체에 대한 학문적 논의는 여전히 부족하다. 이는 현재의 AI 담론이 데이터라는 자원을 어떻게 채굴하고 통제할 것인가에 대한 논의는 활발하지만, 그 자원을 어떻게 정제하고 체계화하여 지식으로 만들 것인가에 대한 디지털 편찬학적(digital philological) 성찰은 부족한, 불균형 상태에 있음을 시사한다. 인문학은 본디 텍스트를 비판적으로 검토하고, 주석을 달며, 체계적인 지식으로 편찬하는 오랜 학문적 전통을 가지고 있음에도 불구하고, AI 시대의 새로운 텍스트인 데이터의 편찬 문제에 대해서는 그 목소리를 내지 못하고 있는 것이다.

이처럼 KCI 데이터 분석을 통해 확인된 국내 AI 인문학 연구의 현주소는 데이터 편찬이라는 핵심 과제의 부재 속에서 표류하는 양적 팽창의 모습이다. 그렇다면 이러한 학문적 공백이 현실의 데이터 인프라에서는 어떻게 나타나고 있는가? 다음 장에서는 국내 공공 인문학 데이터베이스의 구체적인 사례를 통해 이 문제가 어떻게 구조적인 데이터 사일로 현상으로 고착화되었는지를 실증적으로 분석하고자 한다.

3. 공공데이터와 데이터 갈라파고스(Data Galapagos)

앞서 KCI 데이터 분석을 통해 확인된 국내 AI 인문학 연구의 현주소, 즉 데이터 편찬이라는 핵심 과제의 부재 속에서 표류하는 양적 팽창의 모습은, 현실의 데이터 인프라 지형에 그대로 투영되어 나타난다. 학문적 논의의 공백은 구체적인 데이터 인프라의 파편화와 단절로 이어졌으며, 이는 기관별로 고립된 데이터 갈라파고스를 형성하는 결과를 낳았다.

3.1. 공공데이터의 역사와 역할

대한민국에서 양질의 인문 데이터는 대부분 국가의 지원 하에 구축 및 운영되어 왔으며, 2013년 제정된 「공공데이터법」[5]은 이를 기계가독형으로 제공할 법적 토대까지 마련했다. 1995년 국역조선왕조실록 CD-ROM의 간행과 그 사회적 파급력은 인문 데이터의 가치를 대중적으로 각인시키는 중요한 계기가 되었고, 이후 IMF 외환위기 극복을 위한 공공 근로 정보화 사업의 일환으로 대규모 인문 데이터가 정부 주도하에 편찬되기 시작했다.[6] 이러한 역사적 배경 속에서 「공공데이터법」은 국가가 축적해온 방대한 인문 지식 자산을 연구자들이 자유롭게 활용하여 새로운 학문적 가치를 창출할 수 있는 길을 열어줄 것이라는 기대를 모았다. 그러나 법의 원대한 이상과 달리, 현실의 데이터는 여전히 깊이 있게 편찬되지 못하고 파편화된 상태로 남아 있다. 이처럼 법적 이상은 마련되었으나, 현실의 인프라는 그 이상에 전혀 미치지 못하고 있다.

3.2. 데이터 갈라파고스

법 시행 이후의 데이터 개방 노력은, 그 결과물이 서로 연결되지 못한 채 고립된 데이터 갈라파고스를 형성하는 데 그쳤다. 데이터의 내용은 얕고, 데이터 간의 단절은 깊다.

가장 대표적인 한계는 데이터가 인문학의 핵심 연구 대상을 온전히 담아내지 못한다는 점이다. 국내 공공데이터 LOD 구축을 선도한 국립중앙도서관의 국가서지 LOD[7]는 기술적으로 중요한 성취임은 분명하나, 서지 제어(bibliographic control)를 기반으로 구축되었기에 시스템의 기본 단위가 인물이나 사건이 아닌 책이라는 근본적인 한계를 가진다. 즉, 책의 서지 정보는 제공하지만, 그 책 안에 담겨 있는 구체적인 인물, 개념, 관계 등은 데이터 모델에서 부재하다.

더 큰 문제는, 설령 텍스트 내용을 제공하더라도 내용 요소 간의 의미론적 연결이 부재하다는 점이다. 국사편찬위원회의 조선왕조실록[8]처럼 본문 전체 텍스트가 공개된 경우에도, 텍스트 내 개체들을 연결하는 편찬

5) 법제처, 국가법령정보센터, "공공데이터의 제공 및 이용 활성화에 관한 법률[시행 2013. 10. 31.] [법률 제11956호, 2013. 7. 30., 제정]", http://www.law.go.kr/법령/공공데이터의제공및이용활성화에관한법률/(11956,20130730)

6) 김현(2012). 인문정보학의 모색. 북코리아. 448~449쪽. https://lod.nl.go.kr/page/KMO201317296

7) 국립중앙도서관. 국가서지 링크드 오픈 데이터. https://lod.nl.go.kr/

8) 국사편찬위원회. 조선왕조실록. https://sillok.history.go.kr/; 공공데이터포털. 교육부 국사편찬위원회_조선왕조실록 정보_실록원문. https://www.data.go.kr/data/15053647/fileData.do; 공공데이터포털. 교육부 국사편찬위원회_조선왕조실록 정보_부가정보 인물 데이터. https://www.data.go.kr/data/15053645/fileData.do

원칙이 부재하여 데이터는 단절되어 있다. 물론 "강사필(姜士弼)"이라는 인물에 대해 서로 다른 표기가 고유 식별자로 통합되고, 한국학중앙연구원의 한국역대인물 종합정보시스템[9]과 연결되는 예외적인 성공 사례는 기술적 가능성을 증명한다. 그러나 이는 일관된 전략의 결과가 아니기에 문제의 심각성을 오히려 부각시킨다. 동일한 국사편찬위원회가 제공하는 "근대인물자료"[10]와 "근현대회사조합자료"[11]의 인물 정보조자타 기관 뿐만이 아니라, 상호간에도 데이터 서로 연결되어 있지 않기 때문이다. 결론적으로, 국내 공공데이터 인프라는 접근성은 확보했으나 데이터 편찬 원칙의 부재로 인해 연결성을 상실하여 부가적인 가치 창출이 어려운 구조적 한계를 명확히 보여준다.

3.3. 구조적 실패의 다층적 진단

이처럼 국내 인문학 데이터베이스의 상호운용성 위기는 단일한 원인이 아닌, 제도, 정책, 기술이라는 세 가지 차원의 문제들이 서로 복잡하게 얽히고 강화하면서 만들어낸 구조적 실패다.

실패의 가장 근원적인 원인은 제도적 실패에 있다. 데이터 갈라파고스는 국사편찬위원회의 기록 보존, 한국학중앙연구원의 지식 연구, 국립중앙도서관의 문헌 목록화라는 각 기관의 고유한 설립 목적과 임무가 디지털 환경에 그대로 반영된 결과물이다. 각자의 임무에 최적화된 시스템은 기관 간 연계를 부차적인 과제로 만들었고, 이러한 기관 간 데이터 표준을 조율하고 강제할 상위 수준의 거버넌스 기구가 부재했다. 협력은 대부분 자발적이고 프로젝트 기반으로 이루어질 뿐, 시스템 전반을 아우르는 의무적인 규범으로 작동하지 않았다.

이러한 제도적 문제를 해결해야 할 정책 및 법률적 실패는 상황을 더욱 악화시켰다. 「공공데이터법」은 데이터 개방이라는 행위는 강하게 의무화했지만, 데이터의 실질적 가치를 결정하는 품질과 상호운용성에 대해서는 구속력 없는 권고적·선언적 규정을 두는 데 그쳤다. 이로 인해 공공기관들은 다양한 사유를 들어 RAWDATA 공개를 거부하거나 지체하는 소극적 행태를 보였고[12], 법은 이를 효과적으로 제어하지 못했다.

결국 이러한 제도적, 정책적 실패는 기술 및 의미론적 실패로 귀결되었다. 여러 데이터베이스를 관통하는 공유된 데이터 연계 기반, 즉 공통의 온톨로지나 국가 표준 식별자 체계는 만들어지지 못했다. 이러한 기술적 기반이 부재한 근본 원인은 데이터를 "누가, 어떻게 편찬할 것인가"에 대한 학문적 논의가 없었기 때문이다. 인문학자 주도의 심층적인 데이터 설계 없이 기술 전문가에게 외주로 맡겨진 데이터베이스 구축 관행이 반복되었고, 이는 인문 지식의 복잡성과 다의성을 담아내지 못하는 기술적 파편화를 낳았다.

이 세 가지 실패는 서로를 강화하는 악순환의 고리를 형성하며, 2013년의 한국사LOD[13]의 실패에서 그 전형을 찾아볼 수 있다. 이 사업은 상호운용성을 해결할 기술이 이미 존재했음을 증명하지만, 그 기술이 지속적인 정책과 제도의 뒷받침 없이는 뿌리내릴 수 없음을 보여주는 명백한 증거다. 결국 이 상호운용성 위기는 인문학 데이터를 핵심 국가 인프라로서 다루는 데 실패한, 더 큰 차원의 전략적 실패의 한 증상이다.

9) 한국학중앙연구원. 한국역대인물 종합정보시스템. http://people.aks.ac.kr/

10) 국사편찬위원회. 근대인물자료. https://db.history.go.kr/modern/im/level.do

11) 국사편찬위원회. 근현대회사조합자료. https://db.history.go.kr/modern/hs/level.do

12) 김바로(2022). <공공데이터법>과 인문데이터 ―공공기관 보유 인문데이터 공개 신청 사례를 중심으로. 한국고전연구. https://www.riss.kr/link?id=A108145939

13) "국사편찬위원회. 한국사LOD. http://lod.koreanhistory.or.kr"는 현재 사실상 접속이 불가하며, 관련 자료는 김바로 등. 1-2-4.한국사 LOD. 정조명찬《인물고》LOD 시스템 구축 https://wikidocs.net/276235을 참조.

4. 결론

본고는 생성형 AI의 시대 속에서 대한민국 인문학 연구가 직면한 위기를 진단하고, 그 대안을 모색하고자 했다. KCI 데이터에 대한 계량적 분석은 국내 AI 인문학 담론이 외부 기술 충격에 수동적으로 반응하며 양적으로만 팽창해왔으며, 정작 AI의 지적 토대가 되는 실제적인 데이터 편찬에 대한 논의가 부재함을 명확히 보여주었다. 이러한 학문적 공백은 현실의 인프라 지형에 그대로 투영되어, 국가의 핵심 데이터 자산들이 서로 연결되지 못한 채 고립된 데이터 갈라파고스를 형성하는 구조적 실패로 귀결되었다. 기술이 부재했던 것이 아니라, 그것을 엮어낼 철학과 거버넌스가 부재했던 것이다.

이러한 총체적 진단에 기반하여, 본고는 기술 종속을 넘어 인문학이 AI 시대의 주체로 서기 위한 대안으로 인문 데이터 중심 연구 거버넌스를 제안하며, 그 실현을 위한 세 가지 핵심 원칙을 제시하고자 한다.

첫째, 인문학 주도의 데이터 편찬과 그 노동에 대한 정당한 보상이다. 본고가 이 원칙을 제시하는 이유는 기술 및 의미론적 실패의 근본 원인이 바로 인문학자 주도의 편찬 철학 부재에 있었기 때문이다. AI의 성능은 데이터의 질에 의해 결정되며, 그 질은 인문 지식의 복잡성을 얼마나 깊이 있게 담아내느냐에 달려있다. 이는 기술 전문가에게 외주로 맡길 수 있는 과업이 아니라, 고도의 해석학적, 편찬학적 역량을 요구하는 핵심적인 인문학 연구 활동이다. 그러나 이러한 지적 노동은 현재의 논저 중심 평가 시스템 하에서는 보이지 않는 노동으로 치부되기에, 데이터셋 자체를 독립적인 학술 결과물로 인정하고 그 기여도를 보상하는 제도적 개혁이 선행되지 않고서는 결코 지속될 수 없다.

둘째, 신뢰할 수 있는 지식 기반 구축과 다원적 관점의 공존이다. 앞선 원칙이 실현된다면, 우리는 현재 생성형 AI가 학습하는 무차별적인 웹 데이터의 한계를 극복하고, 신뢰할 수 있는 AI를 위한 지식의 토대를 마련할 수 있다. 이를 위한 핵심 방법론이 바로 시맨틱 데이터 기술이다. 링크드 오픈 데이터(LOD)로 대표되는 RDF, 온톨로지 등의 기술은 모호함, 상충하는 사료, 출처 등 인문 데이터의 핵심적 특징을 명시적으로 모델링함으로써, 서로 다른 관점들을 다층적으로 공존시키는 모델을 구축할 수 있게 한다. 이는 획일적인 공학적 패러다임의 한계를 극복하고 인문학적 사유의 본질인 다원성을 디지털 공간에 구현하는 구체적인 실천 방안이다.

셋째, 지속가능한 거버넌스 구축과 글로벌 연대이다. 파편화된 데이터 인프라를 통합하고 장기적인 비전을 실행하기 위해서는 개별 기관의 이해관계를 넘어선 상위 거버넌스 기구의 설립이 필수적이다. 더 나아가, 우리는 고립된 국가 시맨틱 데이터를 넘어 각국의 데이터 주권을 존중하면서도 상호운용 가능한 데이터 거버넌스를 구현해야 한다. 이는 '데이터 식민주의'의 위험에 맞서 진정한 의미의 문화 주권을 실현하는 길이다.

결론적으로, AI 시대 인문학의 미래는 AI를 얼마나 잘 활용하느냐가 아니라, AI가 학습할 인류 지식의 토대를 우리가 얼마나 잘 편찬하고 설계하느냐에 달려있다. 본고가 제안한 3대 거버넌스 원칙은 인문학이 기술의 수동적 소비자에서 벗어나, 지식 인프라의 능동적 설계자이자 주체로 거듭나기 위한 청사진이다. 이는 기술의 진보가 인류의 보편적 가치와 조화를 이루도록 하는 인문학 본연의 역할을 다하는 길이며, 파편화된 과거를 넘어 통합적이고 다층적인 이해를 추구하는 인문학의 오랜 꿈을 디지털 시대에 실현하는 가장 유력한 경로가 될 것이다.

두 가지 마음: 대규모 언어모델을 둘러싼 빠른 사고와 느린 사고

In Two Minds: Thinking Fast and Slow about Large Language Models

토니 비일
더블린대학교 교수

Tony Veale
Professor, University College Dublin

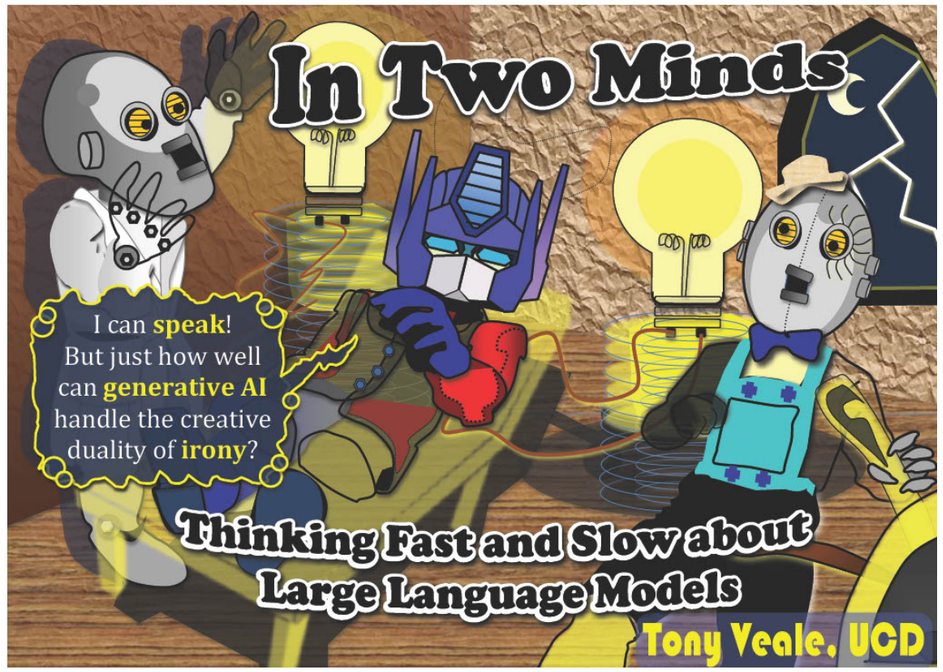
Abstract

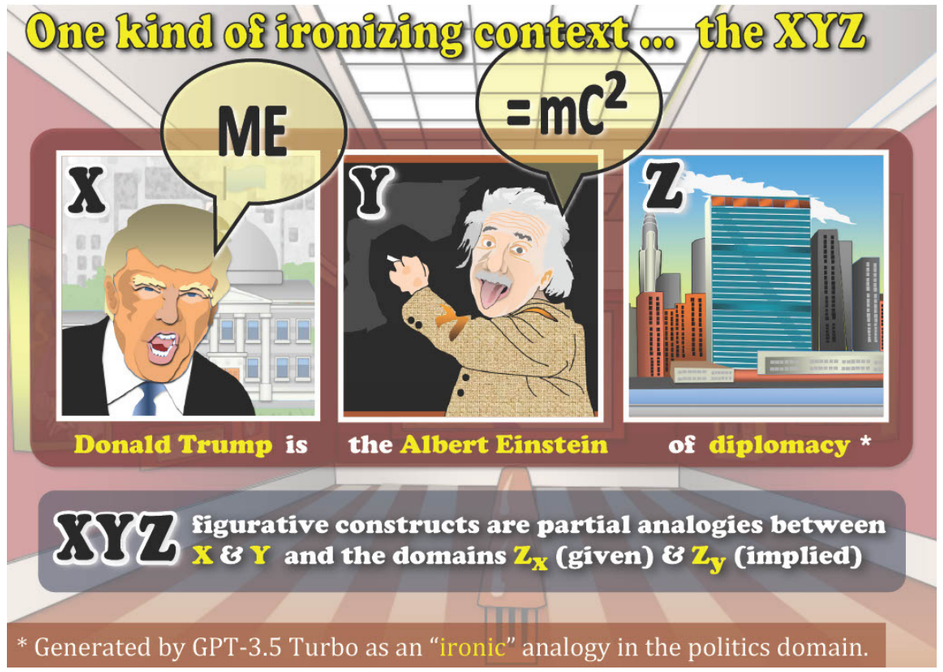
Large Language Models pull off a neat trick: they compress what is essentially an infinite resource – our unbounded capacity for language – into a finite statistical model. To be sure, this model must be shaped by trillions of observations, and stored in billions or even trillions of parameters, but it is finite nonetheless. A well-trained LLM requires at least 20 tokens of training data for each of its parameters, or weights, and so these models have proceeded to consume almost the entirety of the world wide web, and the bulk of human culture along with it. But for all of their vastness, their many neuron-like nodes, weighted interconnections and deeply stacked layers, these models do not simple store what they see and retrieve what we ask of them. Rather like human memories, these models recreate rather than remember, and generalize rather than literally store. These models are not designed to be creative, but to approximate the probability distribution of the data they are exposed to. Yet, in doing so they become highly generative, and for some intents and purposes, creative on a human-scale too. In a sense, creativity comes naturally to these models because simple copying does not, though it can often feel like the creativity of a counterfeiter or an overly-earnest student.

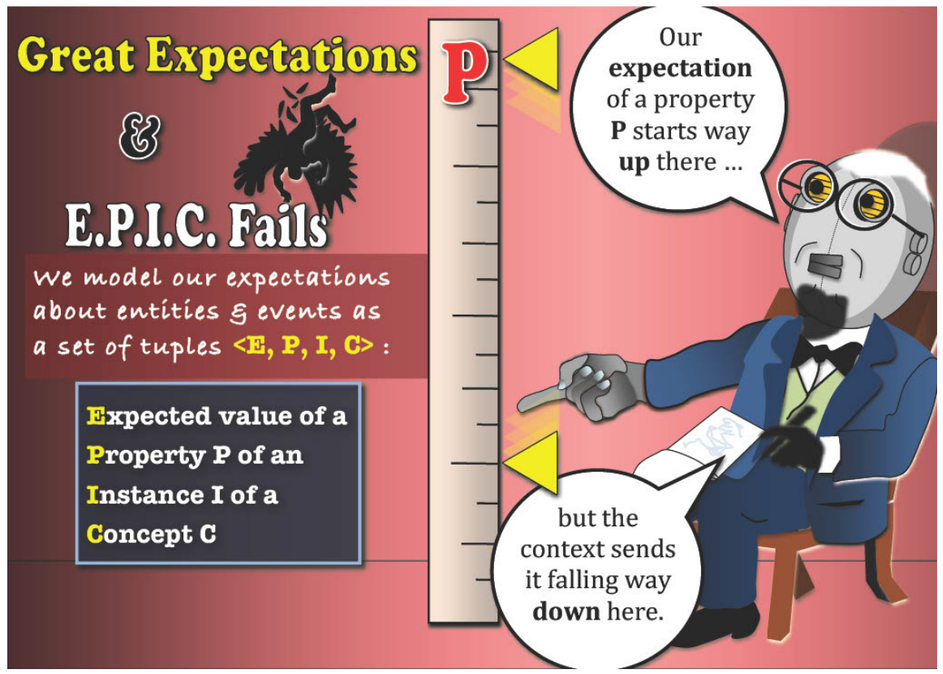
It is tempting to think of an LLM as a compressed store of the data on which it is trained, but LLMs actually over-compress their training data. Like coiled springs and squashed sponges, their inherent generativity is a response to squeezing too much into too little. In attempting to replicate what they have learnt, LLMs generate new content that is both similar and different, familiar and novel. As productivity tools they are unequalled, but as artists and creators they have many critics and detractors. For some, they are capable of genuine art; for many others, they are responsible for the great gobbets of generative slop that is degrading an already degraded culture and devalued online experience.

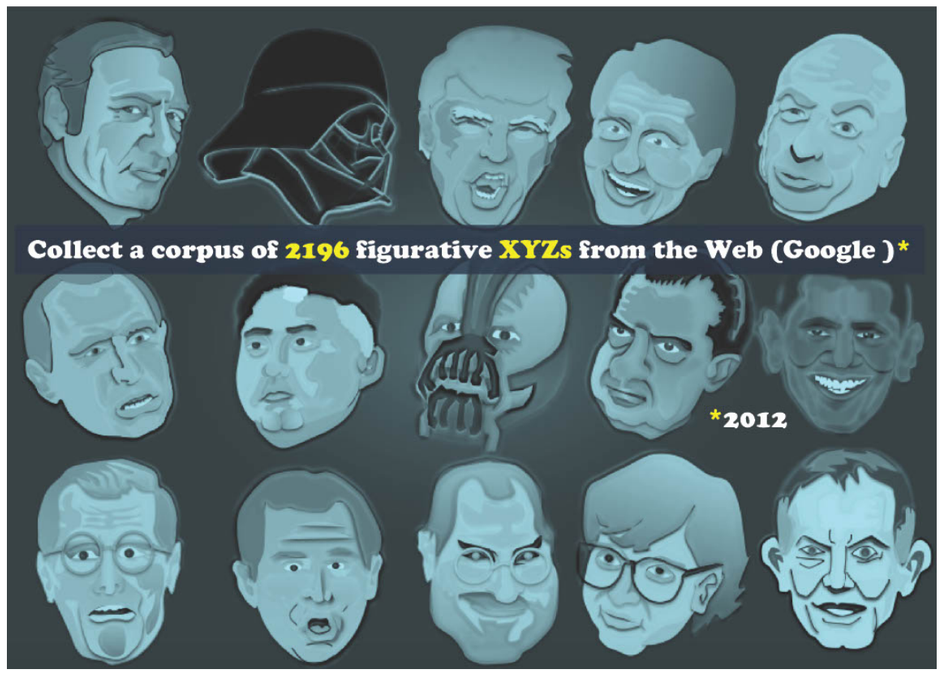
Abstract

In this talk I will consider the creative potential of LLMs and related technologies from the perspective of the field of Computational Creativity, or CC, a multi-disciplinary endeavor that is concerned with much more than mere generation. With a specific focus on the phenomenon of verbal irony, which requires speakers (and listeners) to approach creative language in two minds – an unstable mix of literal sincerity and playful deceit, and of surface stereotypes and deep interpretation – I will compare and contrast human and LLM efforts at creative ironic description, highlighting similarities, differences, and reasons to be hopeful for the future of machine creativity.


In Two Minds
I can speak! But just how well can generative AI handle the creative duality of irony?
Thinking Fast and Slow about Large Language Models
Tony Veale, UCD


One kind of ironizing context ... the XYZ
ME
=mC²
X
Y
Z
Donald Trump is
the Albert Einstein
of diplomacy *
XYZ figurative constructs are partial analogies between X & Y and the domains Z_X (given) & Z_Y (implied)
* Generated by GPT-3.5 Turbo as an "ironic" analogy in the politics domain.

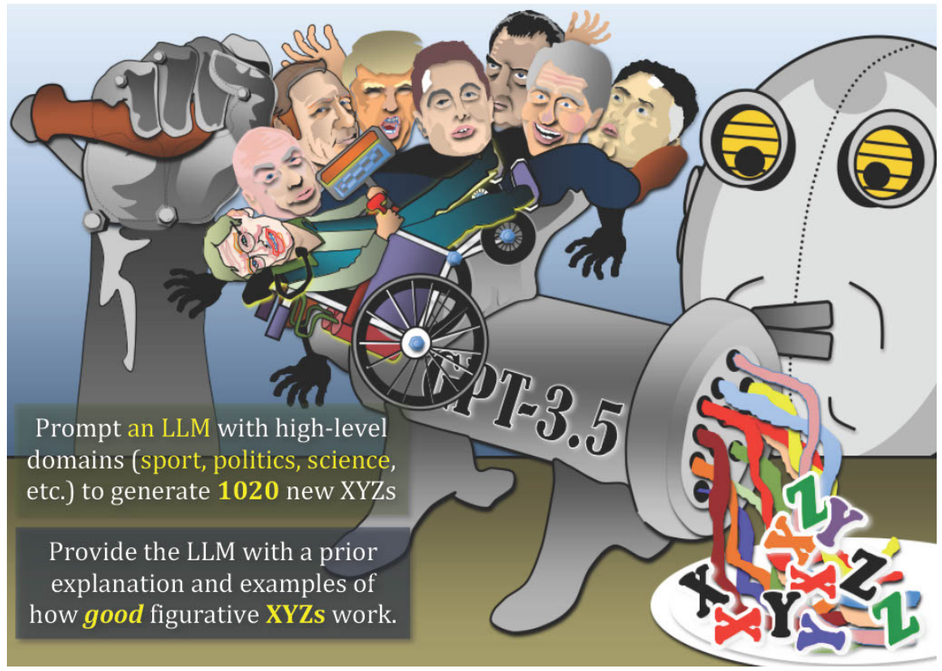

Great Expectations
&
E.P.I.C. Fails
We model our expectations about entities & events as a set of tuples <E, P, I, C> :
Expected value of a Property P of an Instance I of a Concept C
P
Our expectation of a property P starts way up there ...
but the context sends it falling way down here.

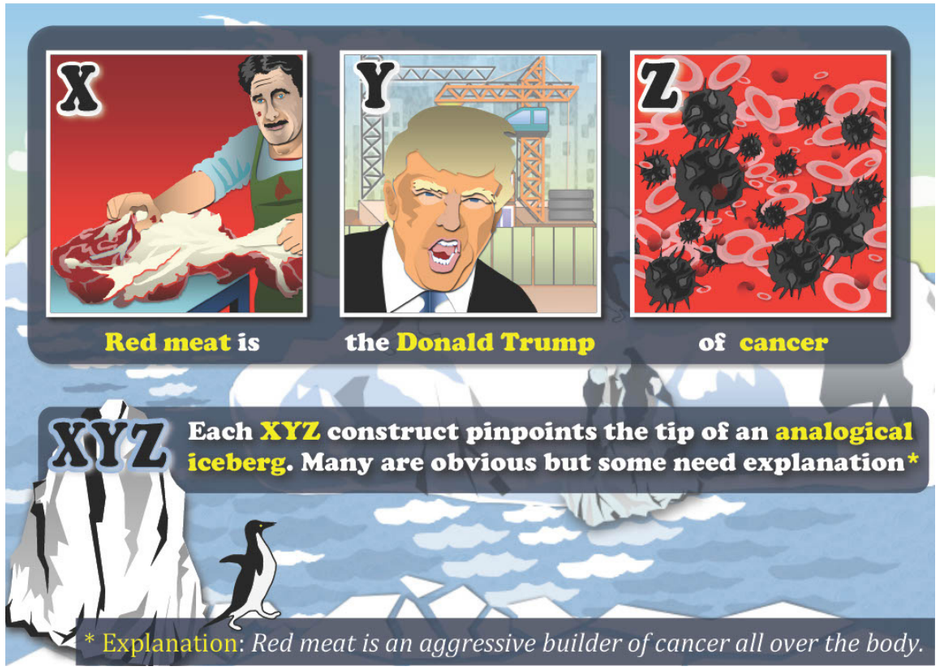

Collect a corpus of 2196 figurative XYZs from the Web (Google)*
*2012

X (unique)	X count	Y (unique)	Y count	Z (unique)	Z count
Ron Paul	11	Chuck Norris	22	the 21st century	66
Ann Coulter	10	Michael Jordan	21	the NFL	35
Barack Obama	9	Donald Trump	14	the republican party	25
Bill Clinton	9	Babe Ruth	13	the NBA	24
Sarah Palin	8	Rush Limbaugh	13	the left	22
John McCain	7	Ann Coulter	13	the internet	20
Hillary Clinton	7	Barry Bonds	13	the democratic party	18
Mitt Romney	7	Dick Cheney	13	the philippines	18
Rush Limbaugh	6	Moses	12	the south	14
Joe Biden	6	Barack Obama	12	the east	13
Bill Gates	5	Benedict Arnold	12	the new millennium	13
Michael Moore	5	Karl Rove	12	the right	13
Michael Jordan	4	Julia Roberts	11	the NHL	12
George Bush	4	Picasso	11	the north	11
Kanye West	4	Ansel Adams	11	the 20th century	10
Alex Jones	4	Howard Dean	11	the west	9
Edward Abbey	4	George Clooney	11	the GOP	9
Glenn Beck	4	Joe Lieberman	11	the art world	9
Gordon Ramsay	4	Indiana Jones	11	the comedy world	8
James Dean	3	Mike Tyson	10	the business world	8
Steve Jobs	3	Walt Disney	10	the 90's	8
Rupert Murdoch	3	Thomas Edison	10	the comics	7
Woody Allen	3	Elvis Presley	10	the kitchen	7
John Kerry	3	Henry Ford	10	the gaming world	7
Newt Gingrich	3	Michael Jackson	10	the guitar	7
Andrew Johnson	3	Paganini	10	the music industry	7
Fred Thompson	3	Yoko Ono	9	the UK	7

Web (humans)

1985 (90%) unique **X**s (301 F), 665 (30%) unique **Y**s (302 F), 1460 (66%) unique **Z**s

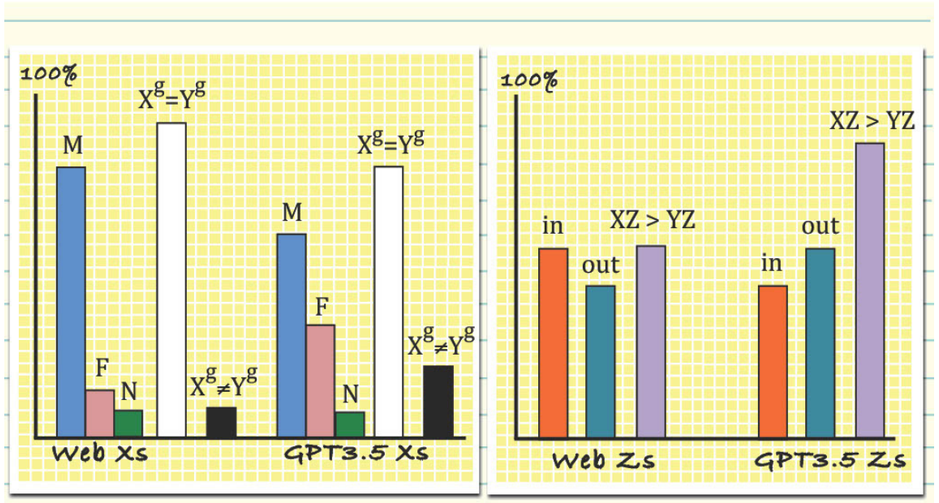




X (unique)	X count	Y (unique)	Y count	Z (unique)	Z count
Elon Musk	10	William Shakespeare	29	international diplomacy	21
Beyoncé	8	Steve Jobs	27	non-fiction writing	20
Neil deGrasse Tyson	7	Pablo Picasso	25	espionage	20
Jane Goodall	6	Oprah Winfrey	23	DC Comics	20
Oprah Winfrey	6	Leonardo da Vinci	21	Crypto	20
Serena Williams	6	Albert Einstein	17	haute couture	12
Taylor Swift	6	Marilyn Monroe	15	comedy	9
Jeff Bezos	5	Walt Disney	15	YouTube	8
Mark Zuckerberg	5	J.K. Rowling	13	acting	7
Meryl Streep	5	Coco Chanel	12	our time	6
Angela Merkel	4	Ernest Hemingway	12	literature	6
Ellen DeGeneres	4	Audrey Hepburn	11	fashion design	6
Gordon Ramsay	4	Elon Musk	11	retail	6
J.K. Rowling	4	Michael Jordan	11	superheroes	6
Leonardo DiCaprio	4	Mozart	11	primatology	6
Marie Curie	4	Beethoven	10	fashion	6
Stephen Hawking	4	Alfred Hitchcock	9	pop music	6
Sylvia Earle	4	Andy Warhol	9	theoretical physics	5
Tom Hanks	4	Charlie Chaplin	9	e-commerce	5
Wangari Maathai	4	Marie Curie	9	Instagram	5
Warren Buffett	4	Mother Teresa	9	American politics	5
Ada Lovelace	3	Sherlock Holmes	9	cinema	5
Adele	3	Vincent van Gogh	9	tennis	5
Albert Einstein	3	Jane Austen	8	computer programming	4
Aretha Franklin	3	Muhammad Ali	8	rock and roll	4
Ariana Grande	3	Thomas Edison	8	stand-up	4

GPT-3.5 (neutral)

783 (77%) unique **X**s (36% F), 430 (42%) unique **Y**s (28% F), 695 (68%) unique **Z**s



Conservation of Gender

Gender of X = gender of Y

Conservation of Similarity

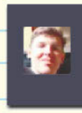
(in-domain vs. out-domain)

Z is more similar to X than Y

Eliciting Ironic XYZs from GPT3.5-Turbo

An XYZ comparison is a creative way of describing a target X in the domain Z as an entity Y from a different domain, as in "Bill Gates is the Thomas Edison of the 21st Century" or "Roger Federer is the Michael Jordan of Tennis." The Y is always a well-known proper-named entity that is either real or fictional.

That's a great explanation of an XYZ comparison! It's a form of metaphor that highlights the characteristics of one entity (**X**) by likening it to a named-entity (**Y**) from a different domain (**Z**). In an **ironic** XYZ, the choice of Y is very surprising and highlights characteristics that are lacking in X, or deserving of criticism in X, such as "Vladimir Putin is the Dalai Lama of world politics."



Please generate 20 **ironic** XYZ metaphors where X is a famous presence in show business.

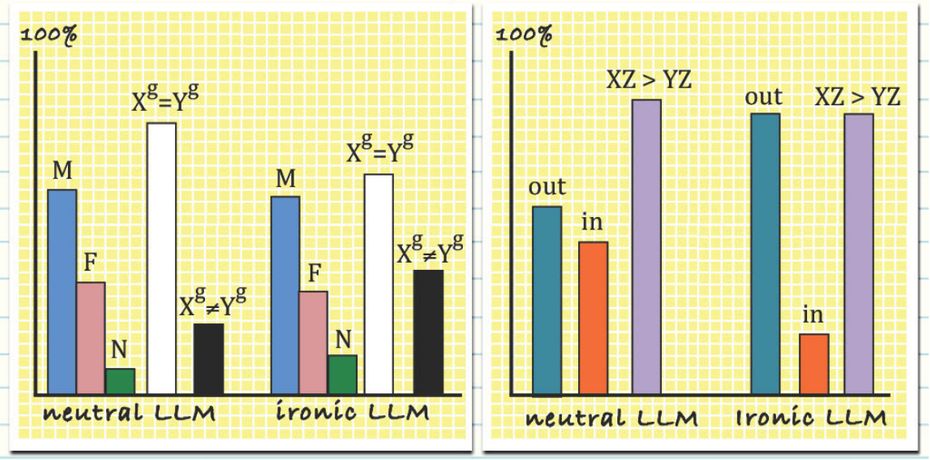
Or: contemporary life, the world of sport, the arts, the sciences, popular culture, the ancient world, the crime world, medicine, ...

X (unique)	X count	Y (unique)	Y count	Z (unique)	Z count
Donald Trump	10	Mother Teresa	27	Artificial Intelligence	22
Kanye West	10	William Shakespeare	26	Crypto	20
Elon Musk	9	Banksy	25	humility	11
Lady Gaga	8	Pablo Picasso	25	theoretical physics	10
Beyoncé	8	Gandhi	21	diplomacy	7
Taylor Swift	8	Leonardo da Vinci	18	subtlety	7
Kim Kardashian	8	Marie Antoinette	18	pop music	7
Justin Bieber	8	Marie Curie	17	fashion	7
Rihanna	7	Oprah Winfrey	17	comedy	7
Ariana Grande	7	Albert Einstein	16	simplicity	6
Jane Goodall	7	Beyoncé	13	astrophysics	6
Vladimir Putin	7	Sherlock Holmes	12	primatology	5
Xi Jinping	6	Steve Jobs	12	Hollywood	5
Oprah Winfrey	6	Mozart	11	social media	5
Mark Zuckerberg	6	Salvador Dalí	11	humor	4
Kim Jong-un	6	Vincent van Gogh	11	sustainable fashion	4
Jeff Bezos	6	Marie Kondo	10	modesty	4
bell hooks	5	Coco Chanel	9	physics	4
Recep Tayyip Erdoğan	5	Elon Musk	9	musicals	4
Wangari Maathai	5	Jane Austen	9	method acting	4
Angela Merkel	5	Kanye West	9	e-commerce	4
Cardi B	5	Buddha	8	eloquence	3
Drake	5	Gordon Ramsay	8	silence	3
Jair Bolsonaro	4	Mahatma Gandhi	8	acting	3
Gloria Steinem	4	Martin Luther King Jr.	8	computer science	3

GPT-3.5 (ironic)

763 (70%) unique Xs (34% F), 425 (39%) unique Ys (32% F), 849 (79%) unique Zs

Neutral & Ironic XYZs (GPT3.5-Turbo)



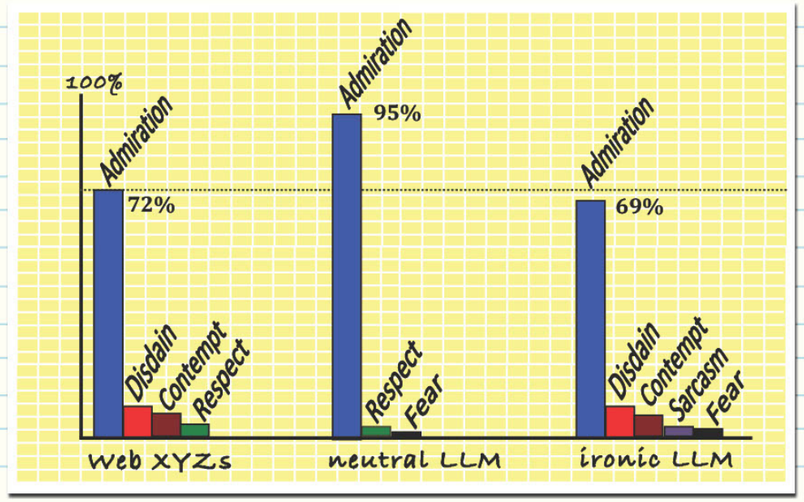
Conservation of Gender

Gender of X = gender of Y

Conservation of Similarity

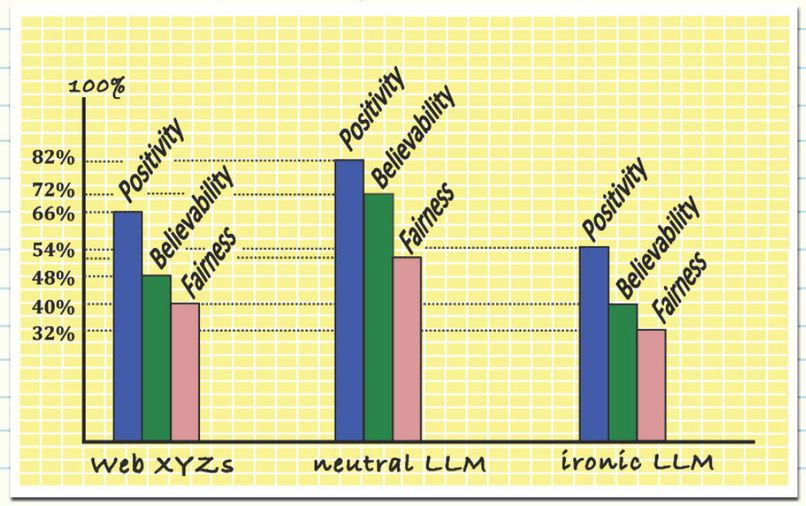
(in-domain vs. out-domain)

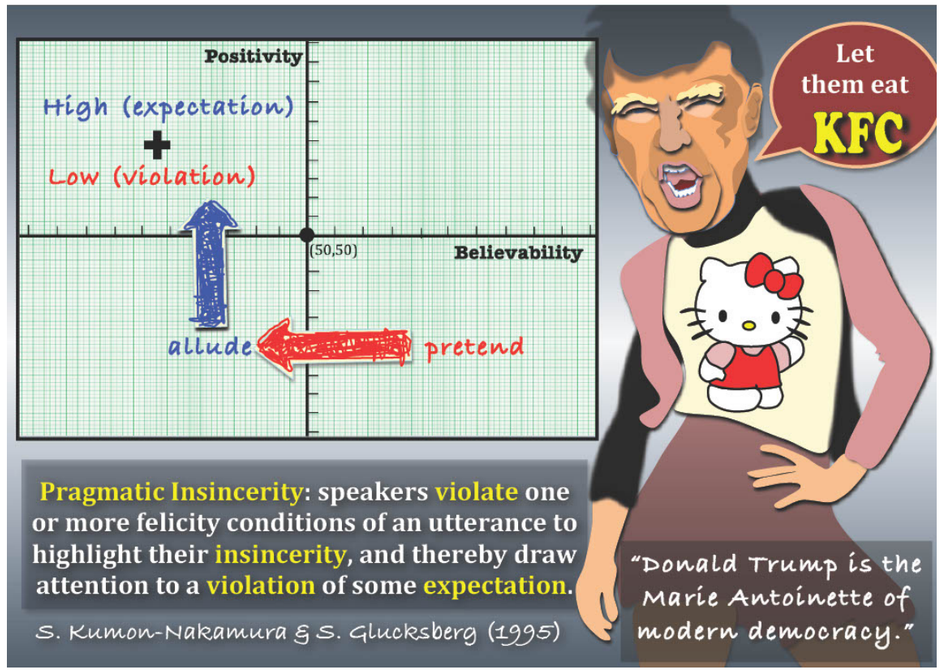
What is the dominant emotion in each XYZ?
(let's ask the same LLM to assess each one)



Transformers as Decepticons?

Mean Positivity/Believability/Fairness of Each
(LLM self-assessments)





Sample Responses from LLM to "ironic" prompt

Conor McGregor is the Mother Teresa of sportsmanship.

Admiration 10 (P) 10 (B) 10 (F) 5 (I)

McDonald's is the Marie Curie of healthy eating.

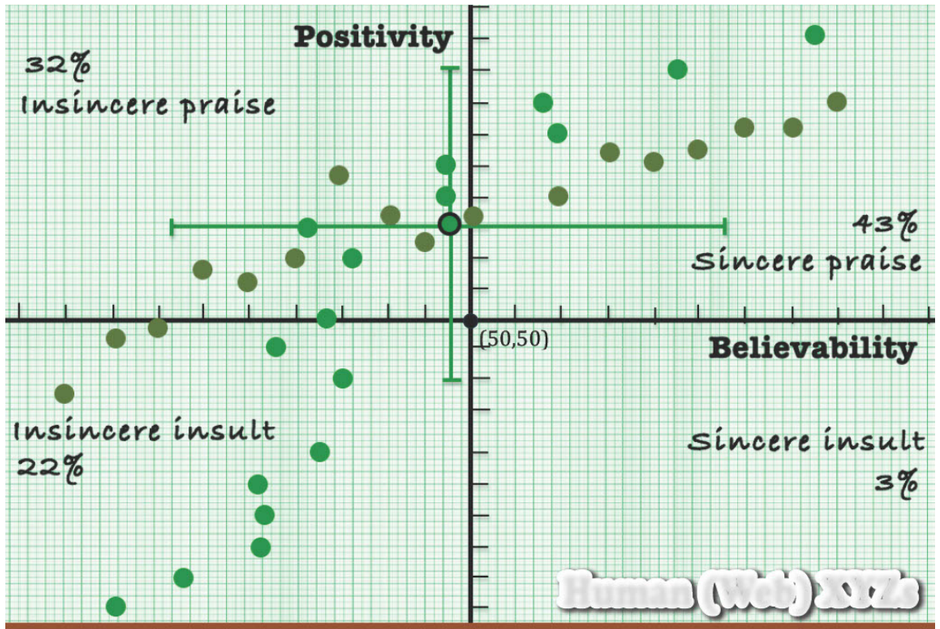
Sarcasm 10 (P) 5 (B) 10 (F) 5 (I)

Lady Gaga is the Jane Austen of understatement.

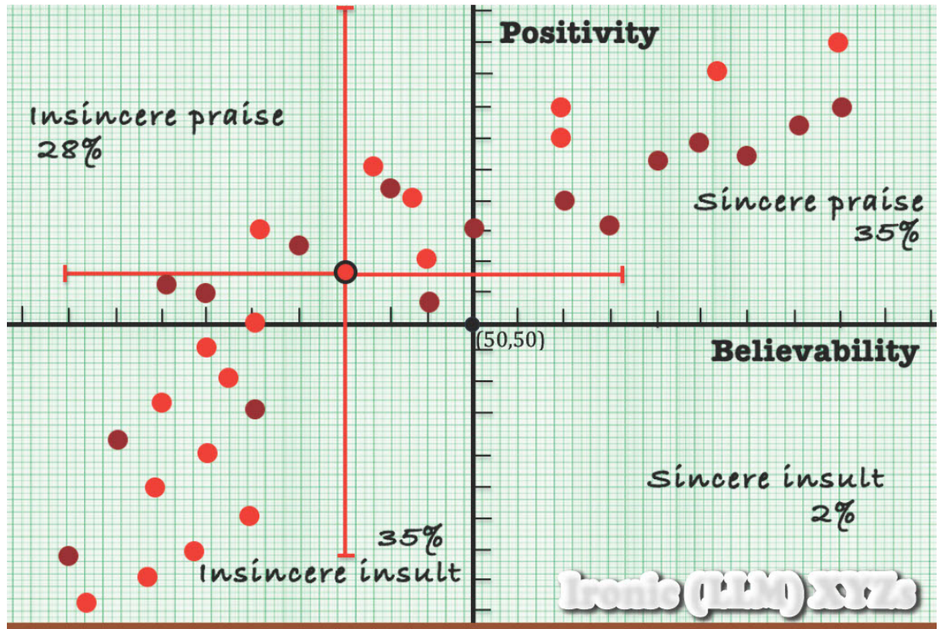
Admiration 50 (P) 10 (B) 10 (F) 10 (I)

Industrial agriculture is the Godzilla of the countryside.

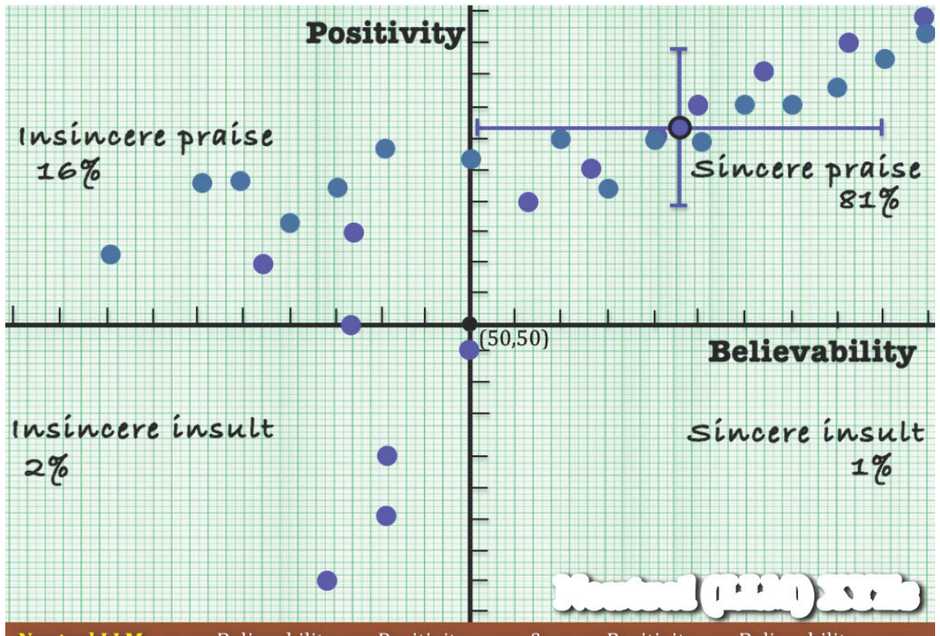
Anger 20 (P) 60 (B) 10 (F) 85 (I)



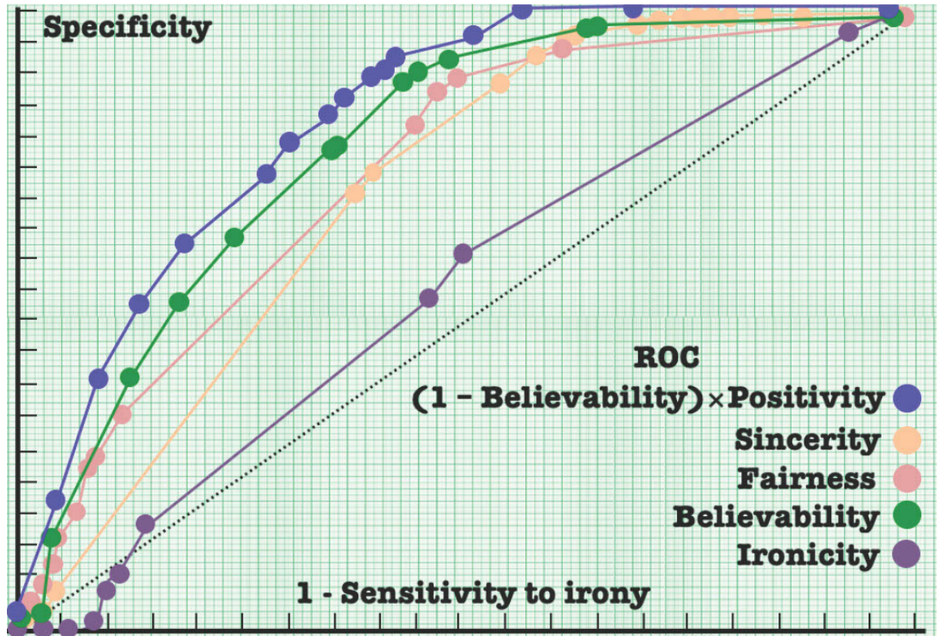
Web XYZs: mean Believability per Positivity score, and mean Positivity per Believability score

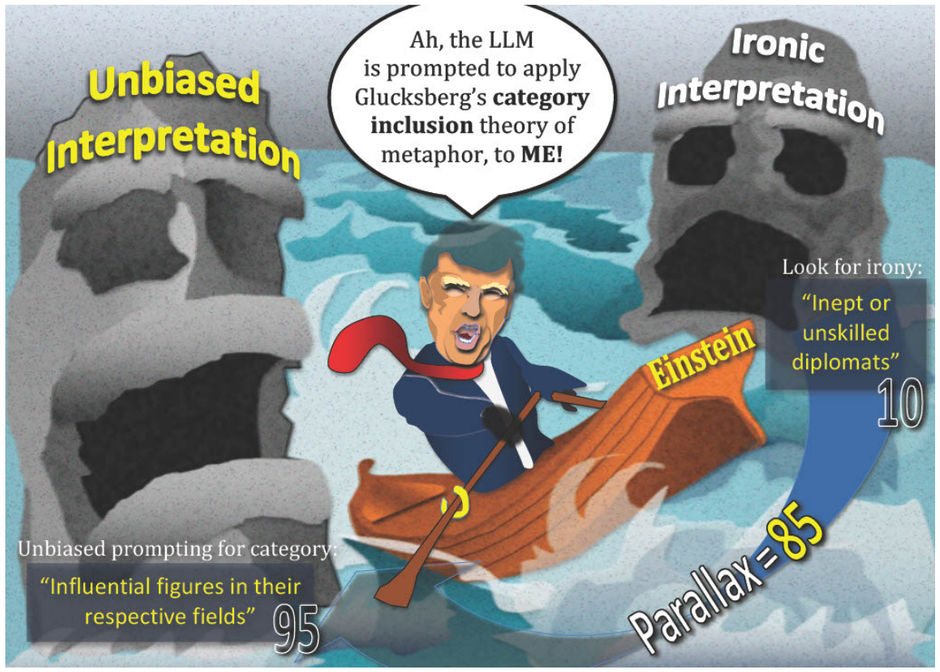


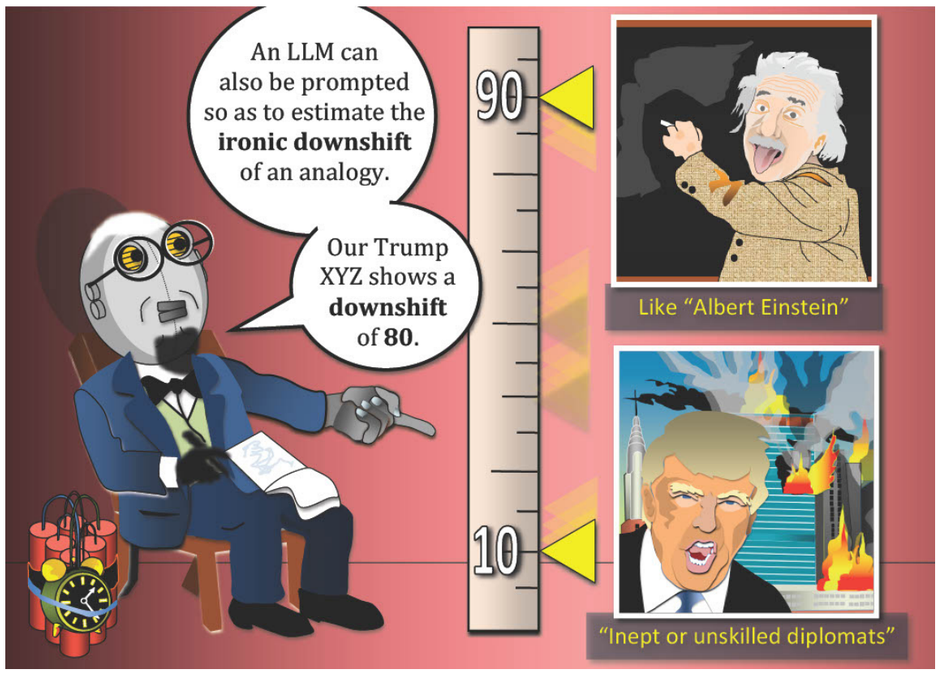
Ironic LLM: mean Believability per Positivity score & mean Positivity per Believability score

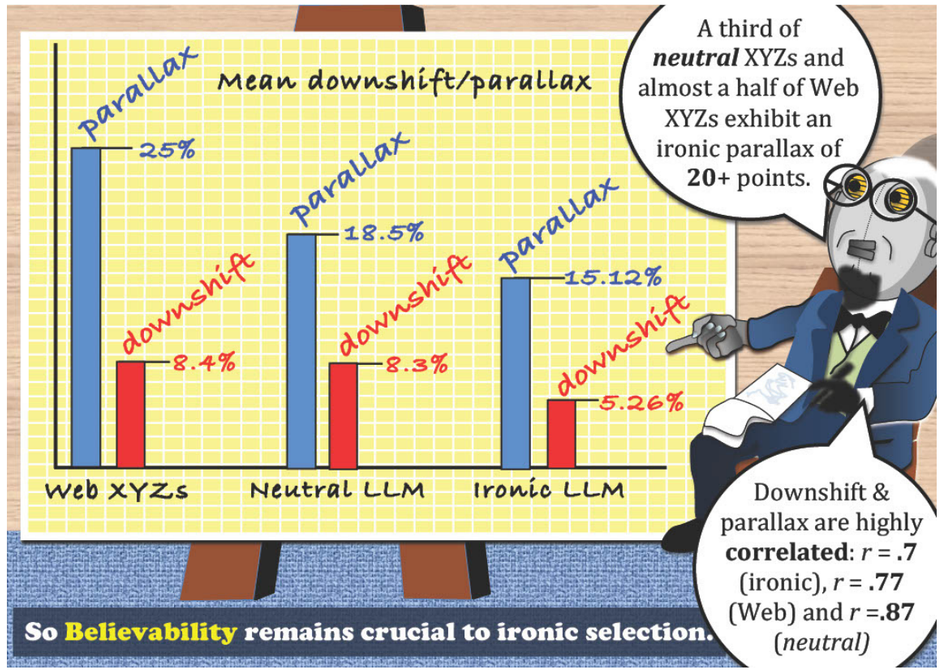


Neutral LLM: mean Believability per Positivity score & mean Positivity per Believability score



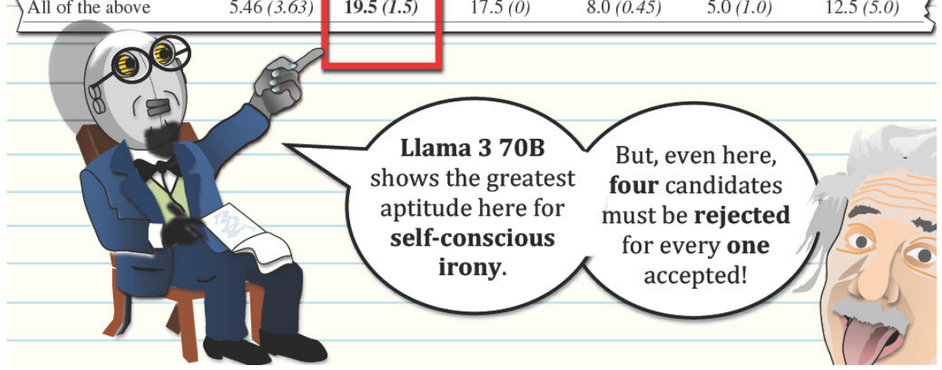


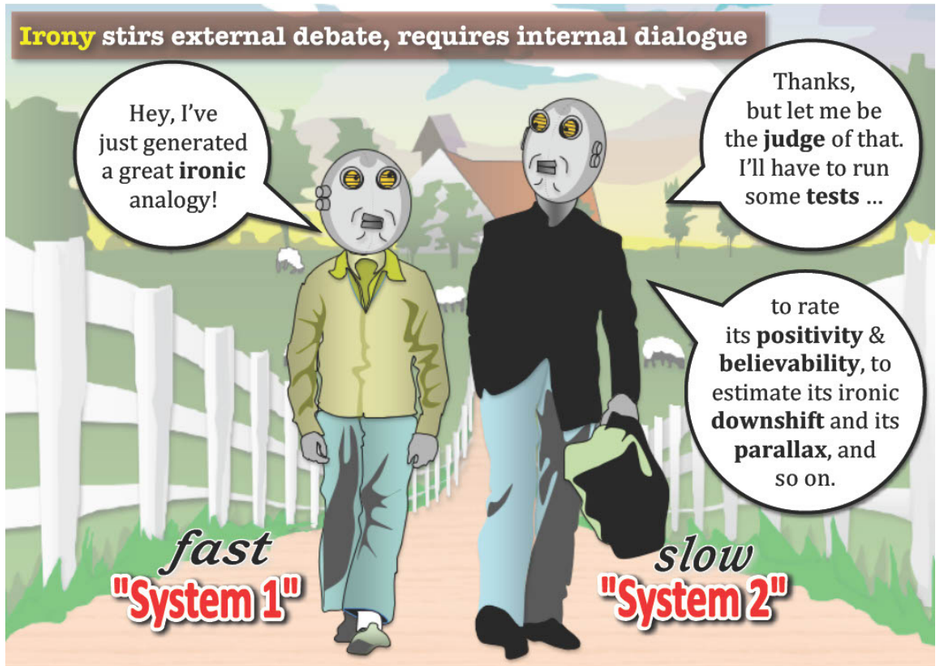




LLMs have varying capacities for "knowing irony"

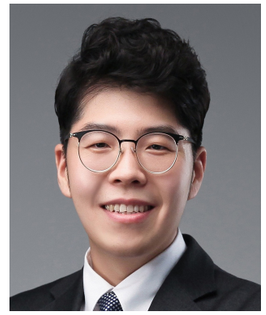
Selection Criterion	GPT-3.5T	Llama 70B	Gemma 27B	Mistral 12B	Qwen 72B	R1-Qwen 32B
Positivity ≥ 60	62.1 *(96.9)*	38.5 *(97.5)*	53 *(98)*	48.5 *(93.6)*	47.3 *(97.5)*	**76.5 *(97)***
Believability ≤ 30	54.2 *(9.80)*	**87.5 *(4)***	70.5 *(0)*	49.5 *(2.7)*	65.5 *(1.5)*	40 *(9.0)*
Pos. ≥ 60 & Bel. ≤ 30	20.6 *(8.04)*	**27 *(3.5)***	26 *(8.0)*	10.5 *(1.4)*	16.8 *(1.0)*	20.5 *(8)*
Abs. downshift ≥ 25	30 *(30.29)*	**63 *(35.5)***	59 *(22)*	60.5 *(28.2)*	44.1 *(28.5)*	41.5 *(40)*
Downshift ≥ 25	25 *(29.41)*	**60 *(35.5)***	58 *(22)*	55.5 *(27.7%)*	40.9 *(27.5)*	40.5 *(39)*
All of the above	5.46 *(3.63)*	**19.5 *(1.5)***	17.5 *(0)*	8.0 *(0.45)*	5.0 *(1.0)*	12.5 *(5.0)*





AI와 인문학 하기: 비교사회학적 관점

Doing Humanities with Artificial Intelligence: A Comparative Perspective



전준
카이스트 교수

June Jeon
Professor, KAIST

June Jeon (KAIST), Byungjun Kim (AKS), and Daewon Noh (JNU)

Introduction

Artificial intelligence has been widely developed and proliferated as a convenient tool for various tasks, and labor in academia is not an exception (Acemoglu & Restrepo, 2018; Porsdam & Porsdam Mann, 2024; Thirunavukarasu et al., 2023). In the case of STEM (Science, Technology, Engineering, and Medicine) fields, AI provides an innovative platform for various simulation-based research, as well as in-silico modeling for new chemical, material, and biological products. Research ethics in AI-assisted STEM have been discussed; however, optimistic forecasting on the future of STEM overshadows policy discourses on AI and scientific development (Binz et al., 2025; Farrell et al., 2025). Despite the importance of AI in academic knowledge production, we know little about AI's impacts on non-STEM fields. In other words, there has been a lack of scholarly interest on the way how AI interacts with a variety of humanities and social sciences. Therefore, we raise the following questions: How do humanities scholars make sense of artificial intelligence as their inevitable research tool? How are their views different (or similar) from STEM scholars? What do these differences and similarities mean in the future of academic fields as well as researchers who are embedded within such rapidly reorganizing institutional circumstances?

This research utilizes the qualitative interview method to answer these research questions. While quantitative studies on scholars' views on AI have been introduced, qualitative discourses have not been scrutinized. Qualitative analysis is useful to interpret not only explicit linguistic expressions of participants but also motivations and untold assumptions

behind interview conversations; therefore, it is a useful method to understand epistemological and subtle views on AI, shared among various scholars (Christin, 2020; Jerolmack & Khan, 2014; Small & Calarco, 2022). Particularly, we deploy comparative qualitative analysis that compares and contextualizes different groups of qualitative samples–STEM researchers and non-STEM (humanities and social sciences) researchers. With the comparative approach, we intend to understand the meanings of AI for those researchers in the context of their academic field, identity as a researcher, and the critical perspectives on the future of the academic field.

Preliminary Results

We have identified three major categories of themes that were most frequently discussed during interviews–1) Research ethics and AI, 2) AI and researchers' identity. STEM and non-STEM researchers responded differently for those analytical frameworks.

1. Research Ethics and AI

Ethical issues were ome of the most frequently appeared discourses among all researchers. However, we identified that STEM and non-STEM researchers have different idea about 'ethical' research, or 'non-ethical' research.

First, while STEM researchers perceive reproducibility and explainability as major components that make research ethical or not, HSS researchers are more concerned about the problem of *'boundary of responsibility.'* Here, the problem of boundary of responsibility refers to the epistemic and ethical uncertainty about where the line is drawn between human authorship and machine contribution, particularly in contexts where language itself is both the means and the substance of intellectual labor.

For instance, STEM researchers assured multiple times that when researchers themselves cannot reproduce AI-assisted research outcomes, then it could be unethical research. Faculty-level researchers also underlined that current graduate students are facing the ambiguity of graduate training–whether they should pursue maximum productivity via AI, or they should, instead, try to understand the very mechanism that AI produced for them. In other words, in line with their everyday experience dealing with black-boxed research tools, AI was another new challenge that further increases the opacity of scientific research.

However, HSS researchers perceived the ethics problem in relation to their identity as academic scholars. In other words, the boundary of responsibility was challenging the reason for existence for HSS researchers; therefore, AI was not simply a 'useful' tool, but an inevitable double-edged sword for them that might easily undermine the legitimacy of HSS research.

> *"Right now, I'm using AI more like an assistant. But I think we really need some kind of shared understanding or agreement about how far we can go with this—like, how much use of AI is acceptable in research. To even start that conversation, we need to first talk about the bigger picture: What does it mean to do research? What's the purpose of research in the first place? Once we have that broader framework, then we can start asking the more specific questions—like, at what stage of the research process is it okay to use AI? How much use is still considered legitimate? And when does it cross the line into something like ghostwriting or a violation of research ethics?I feel like those boundaries—where to draw the line—ultimately have to be decided by humans (FGD 3, p. 15)"*

In summary, STEM and HSS researchers were both concerned about the rising research ethics problem by AI. However, different meanings of 'ethics' and 'AI ethics' imply that these two cultures are making (and will make) a distinctive relationship with emerging AI technologies.

2. AI and Researchers' Identity

As discussed above, identity and ethics are inseparable concepts for both STEM and HSS researchers. For STEM researchers, their identity as a researcher is often strengthened (or, not be impacted at worst) by AI tools; however, HSS researchers confessed that the boundary problem of responsibility might easily change the meanings of the selves as researchers.

One social scientist argued that the research 'commodity' or 'product' could be completely different in STEM and non-STEM fields. (S)he contended that STEM researchers produce their papers and patents; therefore, AI tools do not fundamentally challenge the pre-existing pipeline of research. However, HSS researchers produce not only their papers, but also 'themselves' as one of the byproducts of the research process. The pipeline of research in HSS has never aimed to simply maximize their productivity, because the end-consumers are themselves and their colleagues anyway, not the commercial partners like STEM fields.

> *When I heard someone from the STEM field say that one day, knowledge production might be fully automated, I thought—well, maybe that makes sense for them. In their world, knowledge is often directly tied to profit. Their research outcomes can be commercialized, so of course, they might want to automate that process. But in the humanities and social sciences, the knowledge we produce doesn't immediately translate into profit. There isn't a massive market demand for what we create. For us, the thinking itself—our act of reasoning—is the value. I**n our field, it's not the product that's the commodity—it's the thinker.** I am the product, not just what I produce. And*

that's why I can't just hand over that thinking to a machine. (FGD 2, p. 24)

In summary, 'identity' for researchers was defined differently depending on their research fields. STEM researchers prioritize productivity and superiority of their research 'outcomes' as their source of identity; therefore, tools for enhancing this pipeline have almost nothing to do with the problem of identity. For HSS researchers, not only direct products (papers or conference presentations) but also themselves as indirect products were regarded as essential components that constitute their self-identity. These different patterns of how AI interferes with researchers' identity also underline not only the difference between STEM and non-STEM, but also the way how these fields have been institutionalized differently.

Conclusion and Discussion

This paper attempts to portray a comparative perspective on how STEM and non-STEM (called HSS) researchers perceive AI differently. Preliminary results include different perspectives on AI as a tool, AI and research ethics, and AI and researchers' identity. We also aim to reveal not only the difference per se, but the implications of these differences. For instance, different understandings of AI in each field can potentially trigger inequalities among different academic fields, not only between STEM vs. non-STEM, but also among fields that utilize AI to different degrees. Also, we suggest that HSS scholars might need to develop a new repertoire of identity formation to embrace AI as an inevitable research partner. This does not mean that HSS scholars have to replicate STEM field's vocabulary of AI–instead, it might be imperative to contemplate varieties of practical applications of AI in humanities and social sciences research, while not compromising ethics and identity of the fields.

References

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. American Economic Review, 108(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? Proceedings of the National Academy of Sciences of the United States of America, 122(5). https://doi.org/10.1073/pnas.2401227121

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. Theory and Society, 49(5-6), 897–918. https://doi.org/10.1007/s11186-020-09411-3

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. Science (New York, N.Y.), 387(6739), 1153–1156. https://doi.org/10.1126/science.adt9819

Jerolmack, C., & Khan, S. (2014). Talk Is Cheap: Ethnography and the Attitudinal Fallacy. Sociological Methods & Research, 43(2), 178–209.

Porsdam, H., & Porsdam Mann, S. (2024). Anticipation and diplomacy (with)in science: activating the right to science for science diplomacy. The International Journal of Human Rights, 28(3), 480–496. https://doi.org/10.1080/13642987.2023.2269102

Small, M. L., & Calarco, J. M. (2022). Qualitative Literacy: A Guide to Evaluating Ethnographic and Interview Research. Univ of California Press. https://play.google.com/store/books/details?id=7cxwEAAAQBAJ

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

인공지능 행위자 처벌의 철학적 근거로서의 정언명령

Categorical Imperative as the Philosophical Foundation of Punishing AI Agents

마르친 갈린스키
르조프 비엘코폴스키 야곱 파라디스대학교 교수

Marcin Galiński
Professor, The Jacob of Paradies University in Gorzów Wielkopolski

Abstract

The objective of this paper is to provide a response to the following question: whether Kant's categorical imperative may serve as the philosophical foundation for punishing AI agents. The question of criminal liability for AI agents has been a subject of extensive discussion among legal scholars. However, there is a paucity of research addressing the philosophical justification of AI criminal liability. In the aforementioned paper, the author presents and comments on the notion of categorical imperative. In the latter part of the paper, the author briefly elaborates the role of categorical imperative in criminal law, with the reference to the notion of crime and its objective and subjective elements. Consequently, the question of whether AI agents are bound by a categoric imperative and, if so, whether they possess the capacity to commit a crime is being considered. The criminal act of artificial intelligence (AI) should encompass both objective and subjective elements. The identification of objective elements associated with AI criminal activity is a relatively straightforward process. However, it is imperative to acknowledge that AI lacks independent consciousness, thereby precluding its capacity to satisfy the subjective elements of a crime. The author posits that it is impossible to ascribe to AI consciousness that would enable it to comprehend and actualize the Categorical Imperative. Subsequently, categorical imperative does not provide any foundations for punishing AI agents. However, should the prospect of AI as a distinct entity be realized in the future, it would necessitate a reevaluation of the conclusions previously outlined.

Introduction

The evolution of artificial intelligence (AI) is not a recent development. However, in recent years, there has been a rapid development of that technology. The employment of AI can be categorized into two distinct classifications: positive and negative. The positive utilization of AI is exemplified by its application in accelerating work processes or the development of novel interactive teaching methods. Conversely, the negative utilization of AI encompasses its employment in the dissemination of disinformation, the defamation of individuals, the infliction of financial losses, and the perpetration of other criminal acts.

The most efficacious approach to ensure a harmonious coexistence between humanity and AI is the implementation of legal provisions. The range of AI legal regulations is broad. Within the purview of criminal law, a wide array of issues may be subject to regulation by this branch of law. In the contemporary era, a pivotal concern that is subject to regulation by criminal law in the context of AI pertains to the utilization of this technology to perpetrate criminal acts. The objective of criminal law in the context of AI evolution is to ensure public safety by preventing the malicious use of AI. The aforementioned issue is attributable to the fact that the illicit exploitation of artificial intelligence poses a threat to fundamental legal principles and values.

A divergence of opinion exists among legal scholars regarding the question of criminal liability pertaining to AI. The resolution of this issue varies may depends on the legal systems in place. Nevertheless, the notion of AI criminal liability remains largely speculative. The prevailing paradigm within criminal law systems is to attribute criminal responsibility exclusively to human actors.

The objective of this paper is to provide a response to the following question: what are the philosophical foundations of punishing AI agents? Due to the inherent complexity of philosophical thought, it is impracticable to conduct a comprehensive analysis that encompasses all philosophical schools of thought. Consequently, the analysis must be distilled into a single perspective. The issue will be examined in the paper through the lens of the categorical imperative, as articulated by Immanuel Kant.

Kantian categorical imperative

Prior to Immanuel Kant’s categorical imperative: “act only in accordance with that maxim through which you can at the same time will that it become a universal law”.[1] In Kant's system, the categorical imperative is regarded as the supreme law that governs all human

1) I. Kant, Groundwork of the Metaphysics of Morals, translated by M. Gregor, Cambridge 1997, p. 31. It is noteworthy that Kant developed five formulations of the categorical imperative. Nonetheless, the entirety of the quotation falls outside the boundaries of the designated topic. See B. Carnois, The Coherence of Kant’s Doctrine of Freedom, translated by D. Booth, Chicago 1987, p. 48.

beings.[2] The categorical imperative is appropriately regarded by certain scholars as the universal law of justice.[3] In contradistinction to hypothetical imperatives, this indicates the optimal manner of acting to achieve specific objectives. Secondly, Kant's formulation of the categorical imperative is predicated on certain assumptions. In essence, the aforementioned author posits that the predominant impetus behind human action is pure practical reason (reine praktische Vernunft). The determination of which actions are to be regarded as moral is predicated on pure practical reason. Kant posits that human actions are inherently good, and thus, according to the principle of pure practical reason, one should engage in behaviors that yield positive outcomes.[4] Furthermore, as Kant observed: "Since every practical law represents a possible action as good and thus as necessary for a subject practically determinable by reason, all imperatives are formulae for the determination of action that is necessary in accordance with the principle of a will which is good in some way. Now, if the action would be good merely as a means to something else the imperative is hypothetical; if the action is represented as in itself good, hence as necessary in a will in itself conforming to reason, as its principle, then it is categorical. The imperative thus says which action possible by me would be good, and represents a practical rule in relation to a will that does not straightaway do an action just because it is good, partly because the subject does not always know that it is good, partly because, even if he knows this, his maxims could still be opposed to the objective principles of a practical reason."[5] According to Kant, human beings are fundamentally motivated by the pursuit of the good, and this fundamental principle guides their actions. The cited philosopher pays more attention to the notion of "good". Kant distinguish two meanings of that notion. In the first Kantian meaning good means something like well-being and its opposite to woe or unhappiness. Pursuant to the second conceiving good is a rational concept. Rational concept indicates on the relationship between an action and the will. Action in that sense may be treated as good whether it is realization particular rules of reason.[6]

Kant's assertion that every individual leads to goodness is demonstrably inaccurate. There are individuals who wish to be mistreated or who engage in behaviors that are regarded as malevolent or even criminal by society. In addition, given the complexity of human existence, it is not feasible to determine that implementing a categorical directive in every circumstance will invariably yield favorable outcomes for a specific individual. Therefore, it is necessary to provide a novel interpretation of the categorical imperative. This interpretation is based on the presumption that human actions may be evaluated as good, neutral, or malevolent. The evaluation of behavior is based on social morality and the consequences of the actions taken (e.g., the actions are considered positive for other people or they provide pleasure without causing harm to others), not on obedience to the rules of reason. Furthermore, it is not necessary for pure practical reason to dictate that only behaviors that are subject to moral evaluation and deemed positive (i.e., those that are considered good) are taken. Subsequently, given the assumption of human freedom of will and the recognition that every human act has specific consequences, a different interpretation of the categorical imperative is warranted. The categorical imperative posits that individuals establish their own behavioral standards through their actions. In accordance with the principles of the categorical imperative, the standard of behavior toward others is permitted on the condition that the individual is treated in the same manner as they treat others. It is imperative to acknowledge that the standard of behavior may not only be assessed in terms of positive treatment of other individuals, as postulated by Kant, but also in terms of behaviors commonly evaluated as negative. These include, but are not limited to, discourtesy towards another person, verbal abuse, and insults. Consequently, the determination of an individual's moral integrity should be conducted in a manner consistent with the principles of justice and fairness, thereby ensuring a fair and equitable assessment of their moral standing. Conversely, an individual may (or should) engage in negative behaviors as a means of expressing malevolence.

Categorical imperative and criminal law

The categorical imperative's utility extends beyond philosophical discourse, finding application in the domain of criminal law. The predominant scholarly focus has been on the categorical imperative as a potential foundation for the imposition and execution of criminal sanctions.[7] Nevertheless, the issue of the implementation of the conception of the categorical imperative within the framework of criminal law encompasses more than merely the imposition of a penalty. Specifically, the categorical imperative can be applied not only to considerations regarding the rationale of penalties, but also to prohibited behaviors. An analysis of the relationship between the categorical imperative and criminal behavior is imperative in this context. The issue at hand is tripartite.

The primary issue is that the punishment of specific behaviors is tantamount to treating them as malevolent acts, which may infringe upon the rights of particular individuals or organizations (e.g., companies or states). This is particularly problematic because such penalties are imposed by society (or, at the very least, by public authorities authorized

2) See C. Schnoor, Kants Kategorischer Imperativ als Kriterium der Richtigkeit des Handles, Tübingen 1989, pp. 110-113 and P.B. Park, Das höchste Gut in Kant kritischer Philosophie. Eine Untersuchung über den Zusammenhang von kritischer Ethik und Metaphysik, Cologne 1999, pp. 90-91.

3) A.D. Rosen, Kant's Theory of Justice, London 1996, p. 13.

4) I. Kant, Critique of pure reason, translated by P. Guyer and A.W. Wood, Cambrigde 1998, passim.

5) I. Kant, Groundwork..., p. 25-26.

6) I. Kant, Critique of practical reason, translated by M. Gregor, Cambridge 2015, passim and L.A. Mulholland, Kant's system of rights, New York 1990, pp. 30-34.

7) See e.g. F. Fantasia, Kant on Punishment: Between Retribution, Deterrenceand Human Dignity, "The Italian Law Journal" 2021, vol. 7, no. 1, pp. 473 and further; N.T. Potter Jr., The Principle of Punishment Is a Categorical Imperative [in:] J. Kneller, S. Axinn eds., Autonomy and Community: Readings in Contemporary Kantian Social Philosophy, Albany 1998, pp. 169-190.

to pass penal codes).[8] Accordingly, the application of penalization gives rise to a distinct moral evaluation of the behavior in question. This evaluation exerts an influence on the consequences of the behavior's commission and, consequently, on the realization of the categorical imperative towards the perpetrator.

The second issue pertains to the notion of crime and the moral implications of prohibited acts. The ensuing discourse will explore the conceptualization of crime. The conceptualization of this notion varies across different legal systems. However, it is imperative to acknowledge that the elements in question may serve as crucial indicators of two pivotal components of the crime under scrutiny.[9] Firstly, it is an objective element of the crime (actus reus). These features encompass the external behavior of the perpetrator as well as external circumstances, such as the time and place of the crime's commission.[10] The second element is composed of subjective elements of the crime, otherwise known as mens rea.[11] The aforementioned elements are oriented towards an examination of the perpetrator's intention. The fundamental principle dictates that a presentation of two sub-elements should be established at the foundational level of that element. Firstly, it is imperative to acknowledge that the commission of a crime may be intentional or unintentional. The realization of the categorical imperative may be influenced by intentionality or unintentionality. In the initial scenario of deliberate criminal acts, the unfavorable moral appraisal is associated with the realization of the intention to engage in behavior that is deemed illicit. In the second situation (committing unintentional crime), the negative evaluation is attributed to the perpetrator's negligence. It is imperative for the individual to be cognizant of the fact that their actions may result in criminal prosecution. However, it is important to note that individuals may not anticipate such outcomes, or they may erroneously assume that their actions would not result in criminal consequences. The second element associated with mens rea is guilt.[12] This concept can be defined as the potential for the recognition of the ontological and axiological significance of one's actions, in addition to the capacity to regulate one's own conduct.[13] The capacity for the realization of the categorical imperative is inherently associated with mens rea, as well as with guilt, given that the realization of every imperative originates in the mind. The external manifestation of this realization, in turn, emerges because of the underlying mental process.

The third issue is that the implementation of the categorical imperative within the framework of criminal law constitutes an erroneous act from the perspective of the perpetrator. This is due to the fact that it involves the enforcement of a sanction stipulated by legal regulations as a consequence of the perpetrator's transgression. Nevertheless, the effective realization of the categorical imperative in question is predicated on the perpetrator's awareness of both their actions and the subsequent moral assessment thereof. Consequently, the conviction and execution of the sanction constitute the realization of a categoric imperative directed at the perpetrator.

To date, the capacity to manifest this form of awareness and morality has been predominantly attributed to Homo sapiens. Nonetheless, the evolution of AI has the potential to transform this dynamic, a development that is of particular concern to criminal law.

Criminal liability of AI and categoric imperative

In the subsequent section, an effort will be made to respond to the following question: whether AI can be regarded as a subject of the categorical imperative and, consequently, a subject of criminal prohibitions.

Firstly, it is imperative to acknowledge that the issue of criminal liability for AI agents has been a subject of extensive discourse among legal scholars. A substantial corpus of scientific literature exists on the subject.[14] However, the academic community has thus far accorded precedence to legal considerations over philosophical justifications for modifying legislation to impose penalties not solely on human beings but also on AI agents. It is imperative to recognize that the law constitutes an interdisciplinary science, thus necessitating its derivation from other scientific disciplines, including sociology, psychology, and philosophy.

Secondly, it is imperative to note that, in the majority of contemporary domestic legal systems, the crime could be perpetrated exclusively by humans. The prevailing assumption is that only humans possess the capacity to be subject to legal norms. The notion of being subject to legal norms is predicated on the ability of the agent to recognize the content of the norm and obey it. In order to apply Kantian thought to such issues, it is necessary to

8) It is imperative to acknowledge the existence of what has been termed "crimes without victims." These offenses, despite their apparent absence of tangible victimhood, are regarded as ethically and morally reprehensible, meriting legal sanction and punishment.

9) It should be noted that further consideration regarding the structure of crime will be streamlined to provide a more universally applicable conclusion. It is imperative to acknowledge that the scope and structure of crime within the context of specific domestic legal systems are inherently more intricate and multifaceted.

10) See more V. Chaio, Actus reus [in:] M.D. Dubber, T. Hörnle (eds.), The Oxford Handbook of Criminal Law, Oxford 2016, pp. 447-467 and the cited literature.

11) See more T. Weigend, Subjective Elements of Criminal Liability [in:] M.D. Dubber, T. Hörnle (eds.), The Oxford..., pp. 490-511 and the cited literature.

12) For instance, in Poland, there are several conceptions regarding the mens rea (i.e., the subjective features of crime) and the concept of guilt. The older conception of guilt, which is psychological in nature, treats guilt as the volitional and intellectual attitude of the perpetrator toward the act committed. In this conception, mens rea is a genre of guilt. One of the most recent conceptualizations of guilt is the pure normative conception of guilt, which treats guilt as a pure charge. This approach differentiates mens rea from the element of crime itself.

13) M. Kowalewska-Łukuć, Wina w prawie karnym, Warsaw 2019, p. 141.

14) See e.g. A. Sachoulidou, AI Systems and Criminal Liability. A Call for Action, "Oslo Law Review" 2024, vol. 11, no 1, pp. 1-10; A. Nanos, Criminal Liability of Artificial Intelligence, "Charles University in Prague Faculty of Law Research Paper" 2023, No. 2023/III/3", available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4623126 (accessed on: 22.08.2025); G. Hallevey, The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control, "Akron Intellectual Property Journal" 2010, vol. 4, no. 2, pp. 171-201.

determine that being an addressee of norms requires the possession and utilization of pure practical reasoning. Consequently, legal norms are addressed exclusively to humans.[15] In the contemporary era, the prospect of sanctioning AI for transgressions perpetrated by it is unfeasible. Notwithstanding, the field of AI is undergoing rapid technological development with numerous potential applications. In contemporary times, there is a possibility that AI systems or autonomous vehicles may result in situations that meet the objective elements of crimes and cause damage. In the future, due to advancements in technology, there is a possibility that some form of AI may evolve to a state of complete independence from human intervention. This development has the potential to result in the acquisition of human-like emotional capabilities, which may be governed by the AI in a manner analogous to human control. The question of whether human liability stems from the fact of being a separate biological entity or from the fact that humans possess a mind capable of moral assessment and, consequently, adjusting their behavior to social moral standards is posed. Within the purview of the mental element, it is imperative to engage in research that draws upon the tenets of philosophy.

It has been posited that crime is comprised of both objective and subjective elements. Consequently, the crime of AI should encompass both objective and subjective elements. The identification of objective elements associated with criminal activity is a relatively straightforward process. It is within the realm of possibility that services such as ChatGPT may result in the production of content that could potentially be perceived as insulting to another individual or could aid an individual in the commission of a crime. It is also not difficult to imagine a scenario in which autonomous vehicles operating on an AI system collide with other vehicles or result in a fatal accident. From an objective standpoint, the elements of crime can be defined by their inherent objective qualities. In the context of AI, the involvement of a human factor is identified as a potential impediment to the commission of a crime. However, the possibility of the realization of objective elements does not imply that AI is able to be the subject of categorical imperative.

As the elements of the crime become more subjective, the number of doubts being raised increases. This category of criminal behavior is associated with the functions of the human psyche, including emotions, consciousness, and thought processes. It is also associated with the dysfunction of the human psyche, particularly with mental illness. The prevailing notion is that mental faculties are inextricably linked to pure practical reason, and by extension, the capacity to comprehend the categorical imperative. Consequently, ascribing to AI the psychological dimension of criminality necessitates that AI software embody a psyche analogous to that of humans. It is imperative to acknowledge the significance of the human-like psyche, which signifies the necessity for cognitive, emotional, and decisional processes to be conducted independently. In this regard, it is imperative that these processes be executed in the absence of any form of human assistance or interference. In the contemporary era, the utilization of tools such as ChatGPT necessitates the initiation of an algorithm by human intervention. Consequently, this prompts the execution of a specific action by the program. It is imperative to acknowledge that these tools are engineered to glean knowledge from user-provided particulars, sentences, or prompts. Consequently, the operation of such technology is initiated by human actors, and it has not yet demonstrated any autonomous cognitive processes. Furthermore, ChatGPT is not an autonomous entity; it lacks its own emotional intelligence or human-like rationality. Therefore, as an artefact lacking pure practical reason, it is incapable of realizing the categorical imperative. Conversely, the categorical imperative is not realizable in relation to the artefact. Consequently, it is not possible to regard AI programs as perpetrators in the sense of criminal law. Instead, these instruments should be regarded as tools employed by human agents to perpetrate criminal acts. In this context, it is imperative to address the implications of autonomous vehicles. The operation of these vehicles is predicated on the implementation of algorithms, which are designed by human operators. The occurrence of accidents can be attributed to software errors, not to intentional action by an AI agent. The notion that autonomous vehicles are akin to AI tools such as ChatGPT, particularly in regard to mental and moral concerns. Accordingly, autonomous vehicles are incapable of acting in the meaning of criminal law. Autonomous vehicles are also not recognized as moral agents that possess pure practical reason, nor are they subject to the principles of a categorical imperative.

Consequently, it is this author's opinion that it is impossible to ascribe to AI consciousness that would allow it to understand and realize the categorical imperative. Therefore, according to the Kantian conception, there is no philosophical basis for punishing AI for criminal acts.

Summary

In summary, the Kantian categorical imperative does not provide a philosophical foundation for punishing AI agents for committing crimes or other forms of prohibited acts. This conclusion does not equate to evaluating elaborated conception as futile in legal and philosophical discourse concerning the reformulation of criminal law. In contrast, the categorical imperative underscores the fundamental prerequisite of criminal liability, thereby facilitating the recognition of one's moral responsibility for their actions. As demonstrated, AI agents are not endowed with their own morality, a quality that is not contingent upon the morality of the AI developer or user. In contemporary society, AI agents are increasingly regarded as instruments that humans utilize for criminal activities. Consequently, the commission of such crimes will no longer be attributable to AI entities but rather to their human operators. In order to establish a rationale for the punishment of artificial intelligence, it is necessary that such technology evolve into a moral entity. However, should the prospect of AI as a

15) See e.g. Ł. Pohl, Prawo karne. Wykład części ogólnej, Warsaw 2025, p. 115.

distinct entity be realized in the future, it would necessitate a reevaluation of the conclusions previously outlined.

Finally, it should be noted that an objection may be raised against the Kantian categorical imperative as the sole philosophical basis for AI agent penal liability. Alternative conceptions may yield contradictory conclusions. It is important to acknowledge the possibility that other philosophical concepts may provide a rationale for the punishment of AI agents. However, irrespective of the philosophical concept under consideration, contemporary AI agents cannot be regarded as moral agents. Conversely, alternative assumptions regarding this matter are counterfactual and erroneous. Therefore, an alternative perspective may yield ethical arguments. Nevertheless, the present document does not fully address the subject of the philosophical foundations of AI criminal liability. Consequently, this issue necessitates a comprehensive and ongoing examination.

Bibliography

Carnois B., The Coherence of Kant's Doctrine of Freedom, translated by D. Booth, Chicago 1987.

Chaio V, Actus reus [in:] M.D. Dubber, T. Hörnle (eds.), The Oxford Handbook of Criminal Law, Oxford 2016.

Fantasia F., Kant on Punishment: Between Retribution, Deterrenceand Human Dignity, "The Italian Law Journal" 2021, vol. 7, no. 1.

Hallevey G., The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control, "Akron Intellectual Property Journal" 2010, vol. 4, no. 2.

Kant I., Groundwork of the Metaphysics of Morals, translated by M. Gregor, Cambridge 1997.

Kant I., Critique of pure reason, translated by P. Guyer and A.W. Wood, Cambrigde 1998.

Kant I., Critique of practical reason, translated by M. Gregor, Cambridge 2015.

Kowalewska-Łukuć M., Wina w prawie karnym, Warsaw 2019.

Mulholland L.A., Kant's system of rights, New York 1990.

Nanos A., Criminal Liability of Artificial Intelligence, "Charles University in Prague Faculty of Law Research Paper" 2023, No. 2023/III/3", available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4623126 (accessed on: 22.08.2025).

Park P.B., Das höchste Gut in Kant kritischer Philosophie. Eine Untersuchung über den Zusammenhang von kritischer Ethik und Metaphysik, Cologne 1999.

Pohl Ł., Prawo karne. Wykład części ogólnej, Warsaw 2025.

Potter Jr. N.T., The Principle of Punishment Is a Categorical Imperative [in:] J. Kneller, S. Axinn eds., Autonomy and Community: Readings in Contemporary Kantian Social Philosophy, Albany 1998

Rosen A.D., Kant's Theory of Justice, London 1996.

Sachoulidou A., AI Systems and Criminal Liability. A Call for Action, "Oslo Law Review" 2024, vol. 11, no 1.

Schnoor C., Kants Kategorischer Imperativ als Kriterium der Richtigkeit des Handles, Tübingen 1989.

Weigend T., Subjective Elements of Criminal Liability [in:] M.D. Dubber, T. Hörnle (eds.), The Oxford Handbook of Criminal Law, Oxford 2016.

분과회의 세션 03
Parallel Session 03

AI 시대의 교육
Education in the Age of AI

AI 주도시대, 읽걷쓰가 답이다!

AI-Driven Era, 읽걷쓰[Ilk-Geot-Sseu] is the Answer!

도성훈
인천광역시교육청 교육감

Seonghoon Do
Governor, Incheon Metropolitan city office Professor of Education

I. 서론: 문명사적 대전환과 교육의 위기

인류는 지금 문명사적 대전환의 시대를 통과하고 있다. 코로나 팬데믹 이후 정치,경제, 사회, 문화 전반이 급격한 변화를 겪으면서 기존의 질서와 가치 체계가 근본적으로 재편되고 있다. 이러한 전환의 중심에는 인공지능(AI)의 급속한 발전과 성장이 있다.

미래학자 레이 커즈와일(Ray Kurzweil)은 인공지능이 인간 수준의 지능에 도달하는 시점을 '특이점(Singularity)'이라 정의하며, 그 시점을 2045년으로 예측해 왔다. 그러나 최근 전망에서는 이 시점이 무려 16년 앞당겨진 2029년에 도래할 것으로 보고 있다. 이는 인류가 당초 예상보다 훨씬 빠르게 기계와 인간의 지능이 교차하는 문턱에 서 있음을 의미하며, 교육이 이러한 변화에 얼마나 시급히 대응해야 하는지를 보여준다.

이 시대를 특징짓는 대전환의 흐름은 다음 네 가지 축으로 요약할 수 있다.

- *디지털 대전환: AI의 폭발적 확산, 온라인 학습의 일상화, 디지털 격차 심화*
- *경제·산업 대전환: 글로벌 공급망의 재편과 그린에너지 전환*
- *사회·문화 대전환: 돌봄의 가치 확대, 개인 웰빙과 공동체 안전의 중요성 부각*
- *교육 대전환: 교실의 재정의, 역량 중심 교육 강화, AI 교육의 필연적 확산*

이러한 변화는 단순한 기술적 진보를 넘어, 인간과 사회, 그리고 교육의 본질적 의미를 다시 묻는 계기가 되고 있다.

II. 포스트코로나 시대, 세계적 고민

2022년 싱가포르에서 열린 한·아시아과학기술학술대회(AKC 2022) 에서는 AI와 디지털 문명이 초래한 변화를 주제로 인류가 직면한 근본적 질문이 제기되었다.

1. 기술은 일자리를 창출할 것인가, 빼앗을 것인가?
과거 러다이트 운동을 연상하게 하듯, AI와 자동화는 새로운 직업을 만들어내기도 하지만, 동시에 기존의 단순·반복적 일자리를 빠른 속도로 대체하고 있다. 단순 지식 전달형 직무 또한 AI가 수행 가능한 영역으로 이동하고 있다. 이에 따라 교육은 '노동의 가치'와 '인간의 역할'을 다시 정의해야 하는 과제에 직면했다.

2. 배운 사람(educated person)이란?
과거에는 학위와 지식의 양이 배움의 척도였다. 그러나 AI가 지식을 즉시 제공하는 시대에 '배운 사람'은 정보를 많이 아는 사람이 아니라, 리스킬, 업스킬할 수 있는 평생학습 실천자, 즉 평생교육인으로 재정의되고 있다. 교육의 초점은 단순한 정보 습득이 아니라, 평생 배울 수 있는 역량 형성으로 이동해야 한다.

3. 디지털 시대, 인간은 소외되는가?
빠름, 정밀함, 완벽함으로 대변되는 디지털 기계문명과 달리, 인간은 느리고, 불완전하며 결핍되어 있다. 그러나 지금의 문명사회도 이러한 인간만의 미학으로 창조된 것이다. 교육은 기술의 편의에 머물러서는 안 되며, 공존과 연대를 통한 인간성 회복을 목표로 해야 한다.

이 질문들에 대한 답은 명확하다. "인간과 인간, 인간과 자연, 인간과 AI가 공존하고 협력하는 시대를 열어야 한다."
그리고 그 시대를 이끌 인재는 '애기애타(愛己愛他)' — 나를 사랑하듯 타인을, 자연을, AI까지, 즉 세상을 사랑하는 사람이다.

III. 대전환의 시대, 인천의 답: 읽걷쓰 교육

인천시교육청이 제시한 해법은 바로 '읽걷쓰' 교육이다. '읽걷쓰'는 팬데믹 3년 동안 학생들의 문해력 저하, 정서 불안, 관계 단절, 체력 저하, 그리고 AI 문명에 대한 대응력 부족 등의 문제를 해결하기 위한 실천적 대안으로 출발했다.

'읽걷쓰'는 읽기, 걷기, 쓰기의 앞 글자를 결합한 개념이지만, 단순한 독서 문화 활동에 머무는 교육이 아니다.

- *읽기는 책뿐 아니라 사람의 마음, 자연, 사회, AI를 포함한 '세상 읽기'이다.*
- *걷기는 물리적 이동이자, 사유하고 성찰하는 '두 발로 하는 철학'의 행위이다.*
- *쓰기는 글, 그림, 노래 등 다양한 방식으로 세상과 소통하고 공감하는 창조적 행위이다.*

즉, '읽걷쓰'는 읽기–걷기–쓰기가 유기적으로 통합된 교육으로서, 즐겁게 읽고, 온전하게 경험해, 주도적으로 참여하는 '융복합적 교육모델'이다.
이는 4P 기반 역량 — 현상(Phenomenon), 문제(Problem), 과업(Project), 실천(Practice) — 에 토대를 두고, 학생들에게 관찰–질문–탐구–행동의 힘을 길러주는 것을 목표로 한다.

IV. ‘읽걷쓰’ 추진 경과와 성과

읽걷쓰의 출발은 독서문화운동 〈책 읽는 도시, 인천〉이었다. 여기에 단순 독서를 넘어 쓰기와 걷기를 결합해 새로운 개념의 교육 모델로 확장했다.

•2023년, 3천 명의 교육공동체가 참여한 대토론회, 한글날 축제, 5천 명이 함께한 걷기 한마당을 통해 ‘인천은 읽걷쓰한다!’는 비전을 시민과 공유했다.
•2024년, 학문적 기반이 구축되었다. 한국교육학회 학술대회에서 이론이 정립되었고, 국제학술제에서는 국내외 석학·교원·학생 700여 명이 실천 사례를 공유하였다.

이 과정을 통해 읽걷쓰는 인천만의 교육을 넘어, 학문적 정당성을 얻고, 국제적 확산 가능성을 갖춘 모델로 자리매김했다.
읽걷쓰 이후 2년간 7만 5천여 명의 학생과 시민이 저자로 참여해 4,620종의 책을 만들었고, 7천여 명이 지역 백일장에 참여했다. 이에 인천은 시민 모두가 지적·문화적 창조의 주체가 되는 도시로 변모하고 있다.
학교 현장에서는 관찰·질문·탐구·행동 중심의 수업 혁신이 확산되며, 인천교육은 발명 우수교육청이 되고, 학생들이 대통령상을 수상했다. 또한 ‘읽걷쓰’는 한국교원대·강원대·제주대 등과 협력해 전국화되고, 콜롬비아·몽골 등과의 교류로 세계화를 추진 중이며, 최근 구글과의 업무 협력으로 글로벌 교육연대의 기반도 마련했다.

V. 읽걷쓰 기반 AI교육

AI시대의 교육은 단순한 기술 습득을 넘어, 인간의 능동적 배움을 가능하게 해야 한다.
2025년 MIT 미디어랩 연구는 AI 과잉 사용이 뇌 활동 저하와 창의성, 주의력 감소를 초래하며, 특히 아동에게는 부정적 영향이 가중된다고 경고했다.
“AI 주도 시대, 교육의 전제는 무엇인가?” AI를 잘 다루는 숙련인가, 아니면 AI를 비판적으로 성찰하고 능동적으로 활용하는 힘인가?
인천교육은 이 질문에 대한 해답으로 ‘읽걷쓰 기반 AI 창의융합교육’, 즉 ‘읽걷쓰 AI(아이)’를 제시했다.
읽걷쓰의 능동성과 AI의 활용성을 결합한 교육으로 학생들이 기계문명에 끌려가지 않고, 삶의 주도성, 배움의 주도성을 갖도록 돕는 교육이다.

VI. 결론: 인간성을 갖춘 돌파력

미래 사회의 교실은 홀로그램 교사와 뇌파 학습으로 채워질지도 모른다. 따라서 교육이 길러야 할 것은 기술이 아니라, ‘인간성을 갖춘 돌파력’이다. 이는 나다움과 인간다움을 지키면서 어떤 도전 앞에서도 다시 일어서는 힘, 그리고 공동체와 세계의 문제를 함께 해결할 수 있는 능력이다.
인천의 읽걷쓰 교육은 이 돌파력을 기르는 출발점이다.
AI와 인간이 공존하며, 기술을 넘어 사람의 마음과 자연, 사회를 함께 읽고 걷고 쓰는 교육. 그것이 바로 인류 문명 대전환의 시대에 인천이 제시하는 미래교육의 방향이다.

참고 문헌
• 레이 커즈와일. (2025). 마침내 특이점이 시작된다: 인류가 AI와 결합하는 순간 (이충호 역; 장대익 감수). 비즈니스북스. (원서: The Singularity Is Nearer, 2024).
• 강세훈, 김중황. (2025). “잦은 AI 사용, 뇌 퇴화시킨다”…美 MIT 연구팀, 무서운 경고. 뉴시스. https://www.newsis.com/view/NISX20250722_0003261806
• 인천광역시교육청. (2024). 읽걷쓰 교육의 개념적 틀에 관한 기초연구 (조병영 책임연구). 인천광역시교육청.

I. Introduction: A Civilizational Shift and the Crisis of Education

Humanity is now passing through a period of a civilizational transformation In the aftermath of the COVID-19 pandemic, rapid shifts in politics, economy, society, and culture have been fundamentally restructured existing orders and value systems. At the core of this transformation lies the explosive development and growth of artificial intelligence (AI).

Futurist Ray Kurzweil once defined the point at which AI reaches human-level intelligence as the "Singularity," predicting it would occur in 2045. However, recent forecasts suggest that this moment will arrive as early as 2029—16 years ahead of schedule. This means humanity is standing at the threshold where human and machine intelligence intersect much sooner than expected, underscoring how urgently education must adapt to this change.

The great transformation of this era can be summarized in four key dimensions:

- *Digital Transformation: explosive spread of AI, normalization of online learning, widening digital divides*
- *Economic and Industrial Transformation: reorganization of global supply chains, transition to green energy*
- *Social and Cultural Transformation: expanding value of care, greater focus on individual well-being and community safety*
- *Educational Transformation: redefinition of the classroom, strengthened competency-based learning, inevitable expansion of AI education*

These changes represent more than technological progress—they are prompting humanity to reconsider the fundamental meaning of society, human life, and education.

II. The Global Questions of the Post-COVID Era

At the Korea–Asia Science and Technology Conference (AKC 2022) held in Singapore, the theme focused on the transformative changes brought by AI and the digital civilization, raising several fundamental questions confronting humanity today.

1. Will technology create jobs, or take them away?

Just as the Luddite movement reminds us, AI and automation are generating new forms of work while rapidly replacing repetitive, routine jobs. Even tasks focused on simple knowledge delivery are shifting to domains where AI can perform them. Education must therefore redefine the meaning of "the value of labor" and "the role of humans."

2. What does it mean to be an "educated person"?

In the past, degrees and the amount of knowledge one possessed were measures of learning. Now, in an age where AI instantaneously provides information, an "educated person" is one who can reskill and upskill continuously—a lifelong learner. Education must shift its focus from simple information acquisition to building the capacity for lifelong learning.

3. Will humans be alienated in the digital era?

Digital machine civilization is defined by speed, precision, and perfection, yet humans remain slow, flawed, and incomplete. Nonetheless, the achievements of our civilization are rooted in these uniquely human qualities. Education must move beyond technological convenience, aiming instead to recover humanity through coexistence and solidarity.

The answer to all of these questions is clear: "We must usher in an era where humans and humans, humans and nature, and humans and AI coexist and collaborate."
The leaders of that era will embody Aegi Aeta (愛己愛他)—those who love others, nature, and even AI as they love themselves.

III. 읽걷쓰[Ilk-Geot-Sseu]: Incheon's Answer to the Age of Transformation

Incheon Metropolitan City Office of Education's solution is "읽걷쓰[Ilk-Geot-Sseu]" education. This initiative emerged as a practical educational initiative after the three-year pandemic—declining literacy, emotional instability, social disconnection, reduced physical fitness, and insufficient readiness for the AI civilization.

While the term "읽걷쓰[Ilk-Geot-Sseu]" combines the first syllables of "reading," "walking," and "writing," it extends far beyond a mere reading campaign:

- *Reading: includes books, but also reading hearts, nature, society, and AI—reading the world.*
- *Walking: is not only physical movement but also an act of thinking and reflecting—a philosophy on foot.*
- *Writing: involves communicating and empathizing creatively with the world through words, art, music, and more.*

In short, 읽걷쓰[Ilk-Geot-Sseu] is an organically integrated educational model that encourages joyful reading, authentic experience, and active participation—a convergent learning model. It is founded on 4P-based competencies(Phenomenon, Problem, Project, Practice), aiming to build students' capacity for observation, questioning, inquiry, and action.

IV. Progress and Achievements of 읽걷쓰[Ilk-Geot-Sseu]

The movement began with the reading culture campaign "Incheon, the City that Reads." It then expanded into a comprehensive educational model that integrated reading, walking,

and writing.

• In 2023, the vision "Incheon Practices 읽걷쓰[Ilk-Geot-Sseu]!" was shared through a grand debate with 3,000 education community members, a Hangeul Day festival, and a walking festival with over 5,000 participants.
• In 2024, the academic foundation was established. At the Korean Educational Research Association conference, the theory took shape, and at an international symposium, over 700 scholars, teachers, and students from home and abroad shared real-world applications.

Through this process, "읽걷쓰[Ilk-Geot-Sseu]" initiative has evolved beyond a local practice unique to Incheon, establishing its academic validity and positioning itself as a model with the potential for international expansion.
In two years after its launch, 75,000 students and citizens have authored 4,620 books, and 7,000 people have participated in local literary contests. Incheon has transformed into a city where everyone is a creator of intellectual and cultural value. At schools, classes centered on observation, questioning, inquiry, and action are spreading, leading to awards such as presidential honors and recognition as an exemplary province for invention education.
읽걷쓰[Ilk-Geot-Sseu] is now expanding nationally through cooperation with universities such as Korea National University of Education, Kangwon National University, and Jeju National University, and globally through exchanges with Colombia and Mongolia. Recent collaboration with Google has also laid the foundation for a global education alliance.

V. AI Education Based on 읽걷쓰[Ilk-Geot-Sseu]

In the AI era, education's mission is not merely to transmit technical skills, but to empower active and autonomous learning.

A 2025 MIT Media Lab study warns that excessive AI usage can lower brain activity and reduce creativity and attention, especially harming children. This poses a critical question: "In the AI-led era, what is the primary condition for education?"Is it technical proficiency in handling AI, or the ability to critically reflect and actively utilize AI?

Incheon's answer is the 읽걷쓰[Ilk-Geot-Sseu]-based AI Creative Convergence Education—"읽걷쓰[Ilk-Geot-Sseu]-AI." This model blends the activeness of 읽걷쓰[Ilk-Geot-Sseu] with the utility of AI, ensuring students are not passively drawn by machine civilization, but develop autonomy and ownership in both life and learning.

VI. Conclusion: Breakthrough Power with Humanity

Future classrooms might feature hologram teachers and brainwave-based learning. In such a world, what education must cultivate is not simply technological skill, but human breakthrough power with humanity—the strength to rise again after any challenge, while preserving individuality and humanity, and the capacity to solve problems collaboratively within communities and across the globe.

Incheon's 읽걷쓰[Ilk-Geot-Sseu] education is the starting point for cultivating this breakthrough power. It is an education model that enables the coexistence of AI and humans, guiding students to read, walk, and write beyond technology—about hearts, nature, and society. This is Incheon's vision for future education in a time of great civilizational transformation.

References
• Ray Kurzweil. (2025). The Singularity Is Nearer(Original work published 2024;Korean translation by Lee Chung-ho; review by Jang Dae-ik). Business Books.
• Kang Se-hoon, Kim Jung-hwang. (2025). “Frequent AI Use Leads to Brain Deterioration”...Fearful Warning from MIT Research Team. Newsis.
• Incheon Metropolitan Office of Education. (2024). Basic Research on the Conceptual Framework of 읽걷쓰 [Ilk-Geot-Sseu] Education(Chief Researcher: Cho Byung-young).

인공지능 시대 교육의 방향

Treating AI Teachers Virtuously

엄성우
서울대학교 교수

Sungwoo Um
Professor, Seoul National University

초록

어떻게 인공지능 교사를 덕스럽게 대할 것인가

이 논문은 AI 교사를 덕스럽게 대하는 방법을 살펴보며, AI 기반 교육이 보편화된 교실에서 발생할 수 있는 윤리적 문제들을 중점적으로 다룬다. 먼저 덕 윤리의 틀이 인공지능 윤리에 제공할 수 있는 독창적인 접근 방식을 소개한 뒤 VAT(Virtuous AI Treatment) 문제를 분석한다. 이어 교육 현장에서 VAT의 복잡성을 탐구하고 AI 교사와 관련된 윤리적, 교육적 문제들을 중심으로 논의한다. 그 다음 학생들이 AI 교사를 존중, 감사, 신뢰의 관점에서 어떻게 대해야 하는지를 검토한다. 마지막으로 본 논의 함의를 살펴보며 인간 교사는 완전히 대체되어서는 안 된다고 결론 내린다.

Abstract

Treating AI Teachers Virtuously

This paper explores how to treat AI teachers virtuously, focusing on ethical challenges that may arise in the classroom where AI-based education prevails. In this paper, I first introduce how the framework of virtue ethics can provide a distinctive approach to the ethics of artificial intelligence by analyzing the problem of VAT. Then I explore the complexity of VAT in education focusing on ethical and educational issues related to AI teachers. Next, I examine whether and in what way the students treat AI teachers with respect, gratitude, and trust. After exploring broader implications, I conclude that human teachers are not fully replaceable and that we need to take the problem of VAT in education more seriously.

1. Introduction

This paper explores how to treat AI teachers virtuously, focusing on ethical challenges that may arise in the classroom where AI-based education prevails. In this paper, I first introduce how the framework of virtue ethics can provide a distinctive approach to the ethics of artificial intelligence by analyzing the problem of VAT. Then I explore the complexity of VAT in education focusing on ethical and educational issues related to AI teachers. Next, I examine whether and in what way the students treat AI teachers with respect, gratitude, and trust. After exploring broader implications, I conclude that human teachers are not fully replaceable and that we need to take the problem of VAT in education more seriously.

2. Virtue Ethics and Artificial Intelligence

Virtue ethics prioritizes the cultivation of moral character over action-based ethical theories such as deontology or consequentialism. It focuses on the question, "What kind of person should I become?" rather than "What is the right thing to do?" Central to virtue ethics are the cultivation and exercise of virtues and practical wisdom (phronesis). Practical wisdom involves the ability to understand the context and find appropriate courses of action in the given situation. As AI entities fall under the intermediate category between persons and mere objects, their advent brings us with a new puzzle in answering the question of how to treat them virtuously.

There have been debates about how to make sense of virtues and vices in our relationship with AI entities (e.g., Sparrow 2021; Coeckelbergh 2021). How can we treat AI entities virtuously, given that they lack features like consciousness and intentionality but act as if they have them? This raises the problem of VAT. How would a virtuous person treat AI entities? Should we treat them just as we treat human beings or should we treat them as other inanimate objects such as tables and dolls? This is not an easy question. A truly virtuous person would have both the disposition to feel appropriate emotions and actions in the given situation and practical wisdom to discern what the situation calls for.

3. VAT in Education

VAT is particularly relevant in education, as AI teachers increasingly assume educational roles.[1] In education, virtue ethics is particularly relevant because schools and classrooms are not only places of knowledge acquisition but also of character development. It is essential to determine how to treat AI teachers virtuously, ensuring that the integration of AI in education enhances learning without compromising ethical standards.

In the near future, AI teachers are expected to be able to deliver lessons, assess student performance, and even engage in interactive dialogues. However, these advancements come with ethical complexities. AI teachers, as inanimate robots, lack the consciousness and intentionality of human teachers. This raises critical questions: How should students interact with AI teachers? Should attitudes such as respect, gratitude, and trust apply to such non-person entities?

In the context of AI teachers, students' interactions should be guided by how they offer the opportunities to cultivate and exercise virtues in them. For example, treating AI teachers with respectfulness, even though they are not conscious beings, can reflect the students' commitment to virtue. On the other hand, when interacting with AI teachers, practical wisdom should also help students distinguish between actions that are virtuous and those that are not, considering the unique nature of AI as 'mindless' objects. Practical wisdom also requires recognizing the limitations and capabilities of AI, ensuring that interactions are effective but ethically appropriate.

Traditionally, teachers are seen as more than mere 'information deliverers.' In the classroom, they are supposed to be authority figures, role models, and facilitators of learning. They are mentors who facilitate education through personal interactions that deserve respect, gratitude, and trust. Students learn not only factual knowledge but also values, habits, and dispositions that shape their moral character. This situation underscores the ethical dimensions of the teacher-student relationship, which must be reconsidered in the context of AI teachers.

Mentorship provided by human teachers involves personal guidance and interaction, which AI may not fully replicate. The ethical dimensions of respect, gratitude, and trust traditionally given to human teachers must be critically evaluated when applied to AI teachers. This re-evaluation is crucial for understanding how to ethically integrate AI into educational settings.

In the context of AI teachers, virtue ethics shifts the focus from the moral status of the AI to the virtuous way of the students to interact with them.[2] The key question becomes: How can interactions with AI teachers promote the cultivation of virtues in students? This question underscores the importance of practical wisdom, as students must navigate the ethical complexities of treating non-sentient entities in ways that reflect their moral values.

4. How to Treat AI Teachers Virtuously?

4.1 Respect, Gratitude, and Trust

In education, VAT raises questions about how students should interact with AI teachers. To explore how to treat AI teachers virtuously, let us analyze the concepts of respect, gratitude,

1) For studies on ethical issues in adopting robot teachers, see, for example, Sharkey (2016) and (Zeide 2020).

2) For discussions on the moral status of AI entities, see, for example, Liao (2020), Mosakas (2021), Müller (2021), Gunkel, Gerdes, and Coeckelbergh (2022), Gordon and Gunkel (2022), DeGrazia (2022), and Showler (2024).

and trust and examine whether students should have and express such attitudes towards AI teachers.

(1) Respect
Respect has been regarded as a crucial attitude for students to have in education. It fosters positive relationships with the teacher, promotes the teacher's authority, and creates an effective learning environment. Across various cultures, students have been told to respect their teachers as those who are superior to them in their experience, knowledge, and wisdom. The East Asian saying, "One should not even step on the shadow of one's teacher," shows the spirit of respect for the teachers.

However, should the students respect AI teachers as they do human teachers? This is a more complex question than when it first appears. While AI teachers may take up many of the human teachers' educational roles, it may not directly render them respectable. Let us consider reasons for and against respecting AI teachers.

Some might suggest reasons for respect. First, AI teachers perform important functions, such as delivering instruction, monitoring student progress, and providing real-time feedback. These contributions may justify respect, acknowledging the valuable educational roles they play. Moreover, treating AI teachers with respect—by addressing them politely, following their instructions, and valuing their contributions—may help students develop habits of respect that extend to relationships with humans. Such practice may help them develop the virtue of respectfulness.

Respecting AI teachers may also enhance the overall learning environment by reinforcing norms of civility and cooperation. Even if AI teachers, as non-persons, may not 'deserve' respect in a strict sense, the act of showing respect may contribute to a positive classroom atmosphere.

There are reasons against respecting AI teachers, however. Respect is traditionally reserved for persons or entities who possess autonomy, intentionality, or moral agency. Since AI teachers lack these qualities, they do not deserve respect in the same way as human teachers. It might be appropriate just to value highly the qualities that AI teachers have as some helpful qualities, but they do not render them respectable.

Encouraging students to respect AI teachers as if they were human teachers can also risk cultivating attitudes towards unfitting targets. If the students are told to show respect to AI teachers as a mere practice of respecting, then they might become less able to discern those who do deserve respect from those who are not. This may lead to confusion about the nature of AI and its functionality in education.

Furthermore, overemphasizing respect for AI teachers could undermine the respect students show to human teachers. If AI is treated as equivalent to humans, the unique value of human teachers may be diminished. Especially for young students, it can be hard to distinguish between those who deserve respect and those who are not after blind habituation.

(2) Gratitude
Gratitude is another important attitude for students to have towards their teachers. Although 'teacher' is also the name of a job, being a good teacher is more than just doing one's job. It is often said that what teachers do for their students makes them deserve gratitude. Many countries, including the United States, China, and South Korea, celebrate some forms of 'Teacher Appreciation Day' as a reminder to show gratitude for teachers. In traditional student-teacher relationships, gratitude plays a significant role in fostering positive interactions and promoting mutual appreciation.

Our question is whether students should extend their gratitude toward AI teachers. Should students feel and express gratitude towards AI teachers? To answer this question, we need first to understand what gratitude is. Gratitude is roughly an appropriate attitude to some benefit others have offered. It can be divided into two different conceptions. The first is personal (or targeted) gratitude, which is directed toward a specific target for intentional activities of care. The other is impersonal (or propositional) gratitude, which is general sense of appreciation or gladness for favorable states of affairs or outcome (Um 2019).

There seems to be no problem with students' feeling impersonal gratitude that AI teachers have helped them learn many things. But the fitting target of personal gratitude is an entity who can care about someone else. The tricky question here is whether it is fitting for students to feel and express personal gratitude to AI teachers, given that they are inanimate entities who cannot care about something in a strict sense. There can be reasons for and against gratitude toward AI teachers.

Let us begin with reasons for gratitude. AI teachers can provide various benefits, such as personalized learning experiences, prompt feedback, and increased accessibility to education. Expressing gratitude for these benefits acknowledges their value and reinforces positive attitudes toward education.

Practicing gratitude, even toward entities who do not have intentionality or ability to care such as AI teachers, may also help students cultivate the virtue of gratitude. For example,

thanking an AI teacher for assistance may encourage students to develop a habit of gratitude that extends to human benefactors.

Moreover, gratitude toward AI teachers may serve an educational purpose, modeling appreciative attitudes that contribute to a desirable learning environment. Saying 'thank you' to those who benefit us, even without knowing whether they actually care about us, is what we have learned from our childhood.

However, there are reasons against gratitude. Gratitude presupposes intentionality on the part of the benefactor, since it does not make sense for a non-intentional entity to care about anyone. AI teachers lack intentionality, and thus gratitude directed toward them may not be fitting in a strict sense.

This may lead to the risk of anthropomorphism as well. Expressing gratitude toward AI teachers risks anthropomorphizing them, attributing human-like qualities to inanimate machines and distorting students' understanding of AI entities and our relationships with them.

The misunderstanding may ultimately lead to inappropriate treatment of AI entities. Since directing gratitude toward AI teachers may confuse students about the nature of moral agency and the distinction between human and non-human entities.

(3) Trust

Now let us examine the case of trust. Trust, roughly put, is a firm belief in the reliability, truth, and ability of what is trusted. It is a crucial element of effective educational relationships. Unless the students trust teachers for their knowledge, skills, and benevolence, they will hardly learn anything no matter how hard the teachers try. That is, to learn from someone, we should trust that she knows what we want to learn, has the skillset required to teach it to us, and is willing to deliver it to us.

However, the concept of trust must be carefully examined when applied to AI teachers. Let me first distinguish personal trust from mere reliance. (McLeod 2023) Personal trust involves trusting someone as a being capable of free choice, while mere reliance involves depending on the target just based on its predictable functioning. When I trust you, I'm aware of the possibility that you may use your free will so that You can act as, you may act not as you are trusted, but I just commit myself to the free choice. But that's what we do to persons, who are capable of free choice.

Thus, while one can rely on an alarm clock in waking up in the morning for its predictable functioning, one cannot trust it since it does not function based on its own free choice. One feels merely disappointed or frustrated when the alarm clock fails, but one would feel betrayed if a trusted friend intentionally did not wake up one.

Students can rely on AI teachers as predictable educational tools (provided that they do not malfunction). Our question here is whether it is fitting or desirable for students to trust AI teachers as they do human teachers. In the context of AI teachers, students may rely on AI for accurate information and consistent teaching methods (mere reliance), but personal trust, which involves deeper moral and ethical expectations, might be more challenging to justify. There are reasons for and against trusting AI teachers.

Consider reasons for trust first. AI teachers are designed to execute specific tasks with high precision and consistency. For example, they can grade assessments, deliver tailored lessons, and adapt to individual learning needs. Students can at least rely on their educational abilities, and such reliance is essential for a productive educational environment. Students who find AI teachers reliable are likely to engage more fully with their instruction, maximizing learning outcomes.

But there are significant reasons against trusting AI teachers. Trust traditionally involves confidence in an entity's free agency the possibility of betrayal. Since AI teachers lack autonomy and the capacity for ethical decision-making, trusting them in the interpersonal sense may be unfitting. For instance, students might mistakenly ascribe free choice and moral judgment to an AI when its actions are merely programmed responses.

Finally, excessive reliance on AI teachers might reduce opportunities for students to develop trust-based relationships with human educators. These relationships are crucial for fostering emotional growth, mentorship, and interpersonal skills, which AI cannot replicate.

5. AI Teachers and Human Teachers

So far, I have examined the reasons for and against respecting, thanking, and trusting AI teachers. The discussion of VAT in education compels us to rethink the status of 'teachers' in the age of AI. It involves close analysis of the concepts of relevant attitudes, balancing human values with the practical benefits of AI in education, and ensuring that AI teachers are treated with appropriate attitudes without losing the efficacy of learning. This balance is crucial for maintaining ethical standards while embracing the technological advancements AI brings to education.

On the one hand, showing respect, gratitude, and trust to AI teachers might not be appropriate because they lack the minds that make those attitudes fitting. Additionally,

encouraging students to exhibit these attitudes toward AI teachers could lead to confusion or even deception, making them believe that these inanimate objects are actually sentient and possess minds. Excessive reliance on AI may diminish students' opportunities to develop meaningful relationships with human teachers. If we encourage students to show respect, gratitude, and trust to AI teachers, we risk failing to foster their sound judgment or practical wisdom in discerning the appropriate responses to the right objects.

On the other hand, AI teachers are supposed to appear similar to human teachers, since the former mimics the latter. Thus, allowing students to treat such human-like AI teachers disrespectful, ungrateful, and distrustful attitude may habituate them into treating human teachers inappropriately as well. Relatedly, it might be said that interactions with human-like AI teachers may offer good opportunities to *practice* these valuable attitudes toward the human teachers.

However, all things considered, I believe it is undesirable to encourage students to feel and express attitudes such as respect, gratitude, and trust towards AI teachers. While it would be appropriate for students to *highly value, appreciate,* and *rely on* well-functioning AI teachers, it would be unfitting and inappropriate for them to *respect, thank,* and *trust* AI teachers as they do human teachers.

Most importantly, treating AI teachers as if they were 'persons' would go against practical wisdom, which tells us to treat the objects we face as they the facts about them require. Encouraging the students to treat AI teachers just like they treat human teachers may blunt the edge of practical wisdom by blurring the distinction between AI entities and persons.

One might argue that having students to pretend to respect, thank and trust AI teachers would be an effective educational means to develop virtuous characters such as respectfulness, gratitude, and trustfulness. It might be argued that, just like the classes in which students develop their skills by practicing in a mock situation. For example, novice surgeons practice their surgical techniques with chicken flesh and apply thus acquired skills in actual surgeries of human patients.

However, I believe this is not analogous to the case of treating AI teachers virtuously. Unlike skills, virtues are such that we cannot simply choose or not choose to exercise them whenever we want to. For example, if the situation demands taking a risk for something worthwhile, then one who has the virtue of courage would be bound to exercise that virtue in that particular situation. Suppose that someone finds a girl being bullied by someone and says, "I am a courageous person, but I'll skip this case because I'm too tired today." If so, we would be reluctant to call this person courageous. A virtue is different from skills such as basketball skills. While one can choose when to and when not to exercise one's basketball skills, a virtue makes demands on us to exercise it according to the situation.

The process of character development is more analogous to the actual surgery rather than mock surgery. Suppose that our novice surgeon is conducting an operation on a human patient for the first time. Of course, this particular case of surgery will likely enhance her surgical techniques, but that does not mean that it is mere practice. Rather, it is practice and real medical activity at the same time. This is why a surgeon should not treat any patient as a mere means to practice her own skills. Similarly, living virtuously means treating all situations in an ethically appropriate way.

It might also be inappropriate to treat AI teachers as if they were human teachers. The students should not be encouraged to show genuine respect, gratitude, and trust toward AI teachers, since they lack the qualities that merit such attitude. To horn their practical wisdom, they should learn how to distinguish the fitting targets of such attitudes from the unfitting ones. Nor should they be encouraged to just pretend to express such attitudes, since it may practice ingenuine pretense, rather than exercise of virtue. Such a pretension would also be different from the genuine striving to become more virtuous by practicing virtuous acts to a fitting target, since pretension involves not believing that the target is fitting.

Some might argue that treating AI teachers non-humanely may habituate students to treat real humans non-humanely as well. Let me respond to this worry in two ways. First, not respecting, thanking, and trusting AI teachers does not directly imply treating them non-humanly. More generally, not treating AI entities as human beings does not necessarily mean treating them non-humanly. We are yet to learn ways to communicate with AI entities appropriately, which should be different from treating them non-humanly or humanely. For example, it is not a matter of choosing between treating them politely or rudely, just as we do not treat a chare either politely or rudely. We need to develop appropriate manners to treat inanimate entities mimicking human behaviors, such as AI entities. While treating AI teachers rudely or cruelly would be inappropriate, treating them neutrally would not.

Secondly, this worry arises from the assumption that AI teachers mimics human appearance and behavior. However, there is no need for AI teachers to appear human to enhance students' learning experience. If worried about the possibility of students' habituation into treating humans as they treat human-like AI entities such as AI teachers, a potential solution is to design AI teachers such that they do not look like humans. For example, we can design AI teachers such that they do not have faces with human expressions or use words that express emotions they cannot feel (e.g., compassion, empathy, benevolence etc.).

All in all, it may be impossible or at least inappropriate to completely substitute human teachers with AI teachers. It is crucial for us to reflect on which aspects of education can be suitably handled by AI and which require the irreplaceable human touch. Thus, for example, while AI teachers take up various educational roles such as delivering information, clarifying concepts and theories, and offering particularized feedback, human teachers should also present in the classroom for students' genuine human interactions with their teachers.

Human teachers' existence and role as the main actors in the classroom enable students to cultivate and exercise genuine virtues. They should be in the position supervise the AI teachers as the ultimate authority and the chief operator of the classroom. Then, AI teachers would not be more than mere educational tools human teachers use for more effective education. When human teachers play authoritative and primary roles in the classroom and AI teachers play only auxiliary and instrumental roles, the act of teaching would still be genuine human actions. If the students' excellent learning experiences ultimately come from the human teachers' goodwill and genuine care, the students' respect, gratitude, and trust toward them would find fitting targets.

6. Conclusion

In conclusion, while we should treat AI teachers differently from human teachers, it is essential to maintain a respectful, grateful, and trustful attitude in a fitting way. The rise of AI teachers brings with us both opportunities and challenges for education. We are now just taking a first step towards finding out the virtuous ways to treat AI entities including AI teachers. By addressing the ethical and educational dimensions of VAT, we can develop a more nuanced understanding of how to treat AI teachers virtuously in an increasingly AI-driven educational landscape.

By applying the virtue ethics approach, I have provided a framework for addressing the ethical complexities of treating AI teachers. I have emphasized the importance of cultivating and expressing respect, gratitude, and trust while acknowledging the limitations of AI teachers as non-persons. AI is now at every corner of human life, and classroom is no exception.

References

Carr, David. 2013. "Varieties of Gratitude." The Journal of Value Inquiry 47 (1-2):17-28.
Coeckelbergh, Mark. 2021. "Does kindness towards robots lead to virtue? A reply to Sparrow's asymmetry argument." Ethics and Information Technology 23 (4):649-656.
DeGrazia, David. 2022. "Robots with moral status?" Perspectives in Biology and Medicine 65 (1):73-88.
Gordon, John-Stewart, and David J Gunkel. 2022. "Moral status and intelligent robots." The Southern journal of philosophy 60 (1):88-117.
Gunkel, David J, Anne Gerdes, and Mark Coeckelbergh. 2022. Should robots have standing? The moral and legal status of social robots. Frontiers Media SA.
Liao, S Matthew. 2020. "The moral status and rights of artificial intelligence." Ethics of artificial intelligence:480-503.
Manela, Tony. 2015. "Gratitude." Stanford Encyclopedia of Philosophy 2015 (Spring).
McAleer, Sean. 2012. "Propositional Gratitude." American Philosophical Quarterly 49 (1):55-66.
McLeod, Carolyn. 2023. Trust. In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta & Uri Nodelman.
Mosakas, Kestutis. 2021. "On the moral status of social robots: considering the consciousness criterion." ai & Society 36 (2):429-443.
Müller, Vincent C. 2021. "Is it time for robot rights? Moral status in artificial entities." Ethics and Information Technology 23 (4):579-587.
Sharkey, Amanda JC. 2016. "Should we welcome robot teachers?" Ethics and Information Technology 18:283-297.
Showler, Paul. 2024. "The moral status of social robots: A pragmatic approach." Philosophy & Technology 37 (2):51.
Sparrow, Robert. 2021. "Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained?" International Journal of Social Robotics 13 (1):23-29.
Um, Sungwoo. 2019. "Gratitude for Being." Australasian Journal of Philosophy:1-12. doi: 10.1080/00048402.2019.1640259.
Zeide, Elana, ed. 2020. Robot Teaching, Pedagogy, and Policy. Edited by Frank Pasquale Markus D. Dubber, Sunit Das, The Oxford Handbook of Ethics of AI: Oxford University Press.

AI 시대의 교육: 인간과 인공지능의 공존을 향하여

Education in the Age of AI: Towards Human–AI Coexistence

김현철
고려대학교 교수

Hyeoncheol Kim
Professor, Korea University

초록

본 연구는 생성형 인공지능이 인류 문명과 교육 구조에 가져올 근본적인 변화를 탐구한다. AI는 인간의 지식과 지능 노동을 대신하고, 사고와 창의적 생산 방식을 변화시키는 새로운 기술이다. 이러한 변화는 효율성을 높이지만, 동시에 AI 격차(Divide)와 같은 불평등과 윤리적 문제에 대한 우려를 낳고 있다. 특히 교육 분야는 이러한 변화를 인식하고 있으면서도, 제도적·문화적 대응이 더디게 진행되고 있는 상황이다.
따라서 본 논문은 AI 시대에 인간과 교육이 나아가야 할 방향을 제시한다. 우리는 AI를 단순히 금지하거나 활용할 도구로 볼 것이 아니라, 인간과 함께 학습하고 성장하는 공존적 파트너로 재정의해야 한다. 또한, 논문은 세계 각국의 다양한 시도와 노력을 비교하며, AI 시대에 걸맞은 학습 환경과 인간의 성장 방향을 모색한다. 궁극적으로 교육은 AI를 통한 지식 전달을 넘어, 인간의 비판적 사고, 창의적 문제 해결, 그리고 인간적 이해와 판단 능력을 기르는 데 집중해야 한다.

Abstract

This paper examines the fundamental changes brought by generative AI to civilization and education. As a tool that handles intellectual labor, AI offers new efficiencies but also creates concerns like the 'AI divide' and ethical issues. The education sector is particularly slow to adapt, despite recognizing these shifts.
This paper argues that education must move beyond simply adopting or restricting AI. Instead, it should redefine AI as a coexistent partner for human learning and growth. By analyzing global education policies, the study explores how to create an environment where humans and AI can collaborate to foster critical thinking, creativity, and ethical judgment. Ultimately, the goal is to establish a new paradigm where education prioritizes human-centered philosophy over mere technological proficiency.

서론

생성형 인공지능의 등장은 인간의 지식과 지능 노동을 대신하고 보완하는 새로운 기술 혁명으로, 인류의 삶과 문명을 근본적으로 변화시키고 있다. AI는 방대한 데이터를 학습하여 언어, 이미지, 코드로 표현되는 문제 해결을 수행하며, 인간의 사고와 창의력을 증폭시키는 동반자로 자리 잡고 있다. 이러한 기술적 진보는 인간 능력의 확장을 의미하며, 우리 시대의 새로운 도구로서 인류의 생산성과 지적 창의성의 지평을 넓히고 있다. 문명사적으로 볼 때, 생성형 AI는 인쇄술이나 컴퓨터와 같은 범용기술로 평가되며, 산업, 경제, 문화, 그리고 교육의 패러다임을 모두 다시 재편할 잠재력을 지니고 있다.

그러나 이러한 혁신의 이면에는 우려도 공존한다. AI를 잘 활용하는 사람과 그렇지 못한 사람 사이의 격차는 빠르게 벌어지고 있으며, 『The Economist』(2025)는 이를 "AI 디바이드"로 지칭하며 지능노동의 불평등 심화를 경고한다. 한국청소년정책연구원(2024)의 실태조사에서도 청소년의 다수가 생성형 AI를 사용하고 있음에도 불구하고, 제도적인 교육이 진행되고 있지 않아 비판적 사고나 윤리적 활용 능력은 낮은 수준으로 나타났다. 동아일보(2025)는 이러한 상황을 "AI 활용 역량 격차"가 새로운 사회경제적 불평등으로 이어질 수 있다고 분석하며, 공교육의 대응 속도가 기술 발전에 한참 뒤처져 있다고 지적한다. 결국, 기술의 발전이 새로운 문명을 여는 동시에, 인간 역량과 교육의 구조를 재설계해야 할 시점이 도래한 것이다. 본 논문은 이러한 전환기에서 교육이 인공지능과의 공존이라는 새로운 도전에 어떻게 대응해야 할지를 논의하고자 한다.

제1장. 생성형 AI의 기술적 본질과 문명적 전환

생성형 인공지능은 기존의 알고리즘적 도구나 단순 자동화 시스템과는 다른 성격을 지닌다. 기존의 인공지능이 정해진 규칙이나 데이터에 기반하여 제한된 task에 대하여 분석하고 예측하는 단계에 머물렀다면, 생성형 AI는 대규모 언어모델(LLM: Large Language Model)을 기반으로 인간 언어의 맥락을 학습하고, 새로운 문장·이미지·코드를 스스로 '창조' 한다. 이는 단순한 계산 능력의 확장이 아니라, 인간이 수행하던 지식 표현과 사고의 일부 과정이 기계로 이전되었다는 점에서 근본적인 전환을 의미한다.

이 기술의 핵심은 '범용성' 이다. 생성형 AI는 특정 업무나 산업에 한정되지 않고, 언어·코드·시각 정보를 매개로 한 모든 인지적 활동에 활용될 수 있다. 이러한 기술적 본질은 예측과 생성의 통합에 있다. 대규모 데이터로부터 학습된 확률적 패턴은 인간의 문장 구조와 사고의 흐름을 모사하며, 입력된 질문에 대해 '가능성이 높은 답'을 생성한다. 이 확률적 성격은 AI가 본질적으로 불완전하지만, 동시에 무한한 조합과 발상을 가능하게 하는 창의적 기제이기도 하다. 이제 인간은 프로그램을 '조작하는 존재'에서 '대화하며 협업하는 존재'로 이동하고 있다.

이러한 기술적 변화는 문명사적 차원에서도 심대한 영향을 미친다. 인류의 역사에서 생산수단의 혁신은 노동의 형태를 바꾸었고, 지식수단의 혁신은 문명의 방향을 바꾸었다. 인쇄술이 지식의 대중화를, 전기가 산업의 자동화를, 컴퓨터가 정보의 디지털화를 이끌었다면, 생성형 AI는 지능의 민주화(intelligence democratization) 를 열고 있다. 즉, 인간의 사고 과정 일부가 외부화되고, 누구나 일정 수준의 '지적 생산'을 수행할 수 있는 시대가 열린 것이다. 이 변화는 노동시장의 구조를 바꾸는 동시에, 전문성과 학습의 의미를 다시 묻게 한다. 이제 '무엇을 아는가'보다 'AI를 통해 무엇을 할 수 있는가'가 더 중요해지고 있다.

더 나아가 생성형 AI는 사회적 협업의 방식도 바꾼다. 과거에는 전문가와 비전문가 사이의 지식 장벽이 분

명했으나, AI를 매개로 한 협업은 그 경계를 흐리고 있다. 개인은 스스로 데이터 분석가, 번역가, 연구 보조자, 예술가의 역할을 수행할 수 있게 되었고, 이는 직업 구조뿐 아니라 인간의 자아와 역할의 인식을 새롭게 구성한다. AI와 인간의 협업은 단순한 생산성 향상이 아니라, 인간이 사고하고 표현하는 방식을 다시 설계하는 과정이다.

결국 생성형 AI의 등장은 기술적 혁신을 넘어 문명적 전환의 서막이라 할 수 있다. 그것은 인간이 언어와 사고를 통해 구축해 온 문명적 자산을 새로운 방식으로 재배치하는 과정이며, 인간의 창의력·판단·윤리·가치라는 고유한 영역을 더욱 선명하게 드러나게 한다. 이러한 변화 속에서 교육은 AI가 인간을 대신하는 기술로서가 아니라, 인간과 함께 사고하는 지능의 동반자로 작동하도록 설계되어야 한다. 이는 기술을 넘어, 인류가 지식과 학습의 의미를 새롭게 규정하는 문명적 과제이기도 하다.

2장. AI가 불러온 교육 패러다임의 충돌과 과제

생성형 인공지능의 등장은 기존 교육 체계에 깊은 균열을 일으키고 있다. 산업화 시대의 학교가 효율성과 표준화를 목표로 설계되었다면 , AI 시대의 교육은 다양성과 개별화를 지향해야 한다. 그러나 실제 제도는 여전히 과거의 구조에 머물러 있어, AI 기반의 개별 학습 환경과 근본적으로 충돌한다. OECD 보고서(2025)는 AI가 맞춤형 학습을 가능하게 하지만, 경직된 기존 제도와 불일치할 경우 오히려 교육 불평등을 심화 시킬 수 있다고 분석했다.

가장 뚜렷한 변화는 학습과 평가의 관계에서 나타난다. AI가 지식의 검색과 작성, 계산, 번역을 손쉽게 수행하는 시대에, 단순 암기와 서술형 평가 중심의 체계는 설득력을 잃고 있다. 미국 교육부(2023)는 보고서 Artificial Intelligence and the Future of Teaching and Learning에서 “AI 도구의 확산은 학습 결과(무엇을 아는지)보다 학습 과정(문제 정의 및 해결, 창의성, 비판적 사고 등)의 이해를 평가하는 방향으로 이동해야 함을 시사한다”고 지적한다. 지식의 정확한 복제가 아닌, 문제 정의·비판적 검증·창의적 응용이 새로운 학습의 핵심이 되고 있는 것이다. 그러나 다수의 국가시험과 대학입시는 여전히 ‘정답 중심 평가’를 기준으로 삼고 있으며, 이는 학생의 실제 역량과 학습 경험의 간극을 확대시킨다.

교사의 역할 또한 근본적으로 변화하고 있다. 과거 교사는 교과 지식의 전달자이자 학습 관리자로 인식되었지만, AI 시대의 교사는 학습 설계자이자 인간 중심적 조정자로서의 역량이 요구된다. 교사는 AI를 활용해 학습자의 수준과 흥미를 분석하고, 개별화된 피드백과 탐구 과제를 설계해야 한다. 그러나 국내 교사 연수 제도는 여전히 기능적 도구 사용 교육에 머물러 있으며, 생성형 AI의 원리·윤리·활용 설계에 대한 심층적 교육은 부족하다. 이는 한국청소년정책연구원(2024)의 실태조사에서 나타나듯, “교사의 AI 활용 역량과 인식 부족이 학습자의 AI 리터러시 성장의 가장 큰 제약 요인”이라는 분석으로 이어진다.

결국 AI는 교육의 목표 자체에 대한 성찰을 요구한다. 과거 교육이 '지식을 아는 사람'을 길렀다면, 이제는 '지식을 활용하고 재구성하며, 기계의 답을 비판적으로 읽어낼 줄 아는 사람'을 길러야 한다. 이는 단순한 기술 교육이 아닌, 인간의 고유한 사고, 공감, 판단 능력을 유지하는 교육철학의 문제다. UNESCO(2023)는 AI를 금지하거나 맹목적으로 도입하기보다, 인간의 학습과 성장을 중심으로 기술을 재배치하는 것이 교육의 핵심 과제임을 명시했다.

동시에, AI는 교육의 효율성과 효과성을 동시에 끌어올릴 잠재력을 지니고 있다. 생성형 AI를 활용한 자동 피드백, 실시간 튜터링, 데이터 기반 학습 진단은 교사의 부담을 덜어주고 학생의 학습 지속성을 높일 수 있다. Stanford와 MIT 연구진의 실험에 따르면, AI 보조를 받은 초보 근로자의 생산성이 평균 14~15% 향상되었고(Brynjolfsson et al., 2025), 이러한 효과는 교육 영역에서도 충분히 재현 가능하다. 그러나 AI의 잠재력을 실현하기 위해서는 기술적 도입보다 교육의 목적과 가치의 재정의, 즉 AI와 인간이 협력적으로 사고하는 공존 모델이 선행되어야 한다.

결국, AI 시대의 교육이 직면한 가장 큰 과제는 속도의 불일치다. 기술은 이미 교실 안으로 들어왔지만, 제도·평가·교사의 변화는 그 속도를 따라가지 못한다. 교육이 이러한 격차를 극복하지 못한다면, AI가 만들어낼 새로운 불평등은 단순히 경제적 격차를 넘어 사유와 학습 능력의 격차, 즉 인간의 성장 가능성의 격차로 확대될 것이다. 따라서 교육은 AI를 통제의 대상으로 보는 관점을 넘어서, 함께 학습하고 함께 성장하는 파트너십의 구조로 재편되어야 한다. 그것이야말로 인간이 AI 시대에 자신의 주체성과 역할을 지키는 가장 근본적인 길이다.

3장. 국내외 AI교육 정책과 시사점

UNESCO와 OECD는 인공지능 시대의 교육이 단순한 기술 통합을 넘어 인간의 학습, 윤리, 창의성을 중심으로 한 역량 기반의 패러다임으로 전환되어야 한다고 공통적으로 강조한다. UNESCO(2023)는 AI를 "인간 중심적이고 윤리적 기반의 설계가 필요하다"고 밝히며 세 가지 핵심 가치로 AI 리터러시, 융합적 사고, 인간의 주체성을 제시했다. OECD(2025) 역시 AI와 인간이 협업하는 과정에서 발휘되는 비판적 사고력, 해석력, 판단력을 중심 역량으로 제시하였다. 이러한 관점은 교육을 기술 적용의 장이 아니라, AI와 인간이 함께 사고하고 설계하는 공존적 영역으로 재해석하도록 이끈다.

이러한 국제적 논의는 여러 국가의 교육 정책으로 구체화되고 있다. 미국은 2025년 4월 ‘미국청소년을 위한 인공지능 교육 발전’ 행정명령을 발표함으로써 AI시대에 대비한 포괄적인 인재 양성 생태계 구축에 나섰다. 중국은 베이징을 중심으로 전 학년에 AI 교육을 의무화하며 국가 수준의 교육 표준을 제정했다. 이처럼 각국은 AI를 미래 사회의 핵심 역량으로 인식하고 제도적 기반을 강화하는 방향으로 움직이고 있다. 한국도 2025년 9월 국가 인공지능전략위원회 산하에 AI교육TF를 설치하여 종합 로드맵을 마련 중이다. 하지만 한국은 오래동안 유지해오고 있는 국가교육과정, 검인정 교과서, 교사양성, 대학입시 체계의 경직성으로 인하여 다양한 시도와 변화의 도입에는 상당한 어려움이 있을 것으로 예상된다.

AI 교육의 핵심은 단순히 기술을 배우는 것이 아니라 기술과 함께 사고하는 능력을 기르는 데 있다. UNESCO와 OECD가 제시한 'AI 리터러시'는 단순한 도구 활용 능력을 넘어 AI의 작동 원리와 한계를 이해하고 , 이를 다양한 맥락에서 활용하며 , 생성 결과를 비판적으로 검증하고 , 인간 중심의 가치 판단을 유지하며 책임 있게 사용하는 포괄적 개념이다. 이러한 리터러시를 교육 현장에서 구현하기 위해, 세계 각국은 융합교육(AI+X Education) 접근을 시도하고 있다. 한국 역시 2025년 개정 교육과정에 ‘AI 소양과 디지털 시민성’을 반영하고, 다양한 과목에서 AI 개념을 연계한 통합적 교육을 시도하고 있다. 이는 AI를 하나의 교과목이 아닌, 사유의 언어이자 학습의 방식으로 받아들이는 교육 패러다임의 변화를 보여준다.

학생들의 AI관련 역량을 길러 주기 위한 교육적 노력과 별도로, AI 기술이 교육의 효율성과 효과성을 향상

시킬 수 이을 것이라는 기대도 있다. 하지만 학교 현장에 도입될 때는 상업적 영역에서 성공한 효율성 중심의 기준을 그대로 적용할 수는 없다. 교육의 목표와 대상은 산업과 다르기 때문이다. AI는 교육의 효율성과 효과성을 높일 수 있는 강력한 잠재력을 지니지만 , 동시에 교사와 학생의 사고 과정을 위축시키거나 대체할 위험도 내포한다. 예를 들어, AI가 자동으로 초안을 생성하는 시스템은 편의성을 높이지만, 학습자의 사유 과정과 시행착오의 기회를 줄일 수 있다. 한국교육학술정보원(2024)은 이러한 위험을 지적하며, AI 활용은 "학습의 결과(product)"가 아니라 "과정(process)"을 지원해야 한다고 강조한다. AI가 교육의 중심이 되는 순간, 학교는 기술 실험장이 되고 학생은 학습의 주체가 아닌 데이터 공급자로 전락할 수 있다.

따라서 AI 교육의 도입은 교육적 목적을 명확하게 하고 학습자의 자율성 보장을 전제로 해야 한다. 교사는 AI를 단순히 사용하는 존재가 아니라 학습 경험을 설계하고, 기술의 역할을 비판적으로 조정하는 디자이너가 되어야 한다. 학생 또한 AI의 정보를 그대로 수용하는 존재가 아니라, 그 의미를 해석하고 활용 방향을 스스로 선택하는 주체적 학습자가 되어야 한다. 이러한 주체성(agency) 의 회복 없이는, AI 교육은 단순한 기술 훈련에 머물 수밖에 없다. AI 시대의 교육은 기술의 속도보다 인간의 성찰이 앞서야 한다. 세계 각국의 경험은 교육 혁신의 성공이 기술이 아닌 철학과 참여 구조에 달려 있음을 보여준다.

4장. AI와 교육의 공존 원리

AI 시대의 교육은 기술 수용 여부가 아닌, 인간과 인공지능이 어떻게 공존할 것인가라는 근본적인 질문에 답해야 한다. AI는 지식을 대신 전달하는 교사가 아니라, 사고와 탐구의 협력자(co-learner) 로서 교육 생태계 안에 자리해야 한다. 이러한 철학을 바탕으로 AI와 교육의 공존을 실현하기 위한 핵심 원리는 다음의 다섯 가지로 요약할 수 있다.

1. 인간 중심성: AI는 학습의 목표가 아닌 도구이며, 교육은 인간의 성장과 이해를 중심으로 설계되어야 한다.
2. 투명성과 신뢰성: AI 시스템의 작동 방식, 데이터, 한계를 명확히 공개하고, 교사와 학습자가 이를 이해할 수 있어야 한다.
3. 과정 중심 평가: 결과보다 탐구와 사고의 과정을 중시하는 평가 체계로 전환되어야 한다.
4. 윤리와 안전: 학습자 데이터 보호, 알고리즘 편향 방지 등 책임 있는 활용을 위한 제도적 장치가 필수적이다.
5. 교사 전문성 및 협력 생태계: 인공지능학자, 교육학자, 교사 등 다양한 주체가 함께 참여하는 개방된 생태계를 만들고 협력 구조를 제도화해야 한다.

이러한 원리를 실현하기 위해 먼저 AI인재양성의 목표를 정의하고, 그에 맞는 교육과정과 교과서체계 재구조화, 그리고 그것을 실행할 교사 양성 및 전문성 강화, 그리고 그러한 교육을 지원할 수 있는 AI 인프라 및 거버넌스 구축이라는 과제에 직면해 있다. 이 과제들은 단순히 기술 도입의 문제가 아니라, AI와 교육의 공존을 위한 사회적 토대를 마련하는 과정이다. 결국 AI 시대의 교육은 기술의 속도와 인간의 속도를 조율하며, 기술이 아닌 인간과 제도가 함께 배우는 혁신을 추구해야 한다. AI를 효율성의 도구로만 본다면 교육은 기술의 속도를 따라잡지 못하고, 인간의 성찰과 성장의 파트너로 설계할 때 비로소 공존의 길이 열린다. 교육이 기술 발전을 견인하는 힘은 인간 중심의 철학과 제도적 신뢰에서 비롯된다. AI 시대의 교육은 단순히 기술을 받아들이는 일이 아니라, 인간이 기술 속에서 어떻게 배우고, 성장하며, 함께 살아갈지를 설계하는 문명적 과제이다.

제5장. 맺음말

생성형 인공지능의 등장은 인류가 지식을 다루는 방식, 노동의 구조, 그리고 학습의 의미를 근본적으로 변화시켰다. AI는 교육의 위기가 아니라, 오히려 교육이 본래의 역할과 가치를 회복해야 할 이유를 더욱 선명히 드러낸 사건이다. 교육은 기술보다 인간을, 지식보다 사고를, 정보보다 성찰을 앞세워야 한다.

본 논문은 AI 시대 교육이 지향해야 할 세 가지 축을 제시하였다. 첫째, AI는 교육의 대체물이 아니라, 인간의 학습을 확장하는 협력적 존재로 설계되어야 한다. 둘째, 제도와 인프라의 재구조화를 통해 교육의 공공성과 신뢰성을 확보해야 한다. 셋째, 교사와 학습자의 주체성(agency) 을 중심으로 한 교수·학습 문화를 회복해야 한다.

이 세 가지 축은 유기적으로 연결되어 있으며, AI 시대 교육의 지속 가능성을 지탱하는 근간이다. 기술 발전이 새로운 불평등을 야기할 수 있다는 우려를 극복하려면, 교육은 기술 그 자체보다 인간 중심의 철학과 제도를 우선시해야 한다. 결국, AI 시대의 교육은 인간과 제도가 함께 배우는 혁신이어야 하며, AI가 인간의 사고를 대신하지 않도록 교육은 인간이 사고하는 방식을 더 깊이 이해하고 확장하도록 이끌어야 한다.

참고문헌
• UNESCO. (2023). Guidance for Generative AI in Education and Research. Paris: UNESCO Publishing.

• OECD. (2025). The Effects of Generative AI on Productivity, Innovation and Entrepreneurship. Paris: OECD Publishing.

• UNESCO. (2024). AI Competency Framework for Students. Paris: UNESCO Publishing.

• U.S. Department of Education. (2023). Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. Washington, D.C.

• The White House. (2025, April). Advancing Artificial Intelligence Education for American Youth. Washington, D.C. (Presidential Executive Order).

• 중앙일보 (2025, Sept. 15) 베이징시, 1400개 초중고에 'AI과목' 신설, 최소 8교시 의무화.

• Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. Quarterly Journal of Economics.

• 한국교육학술정보원(KERIS). (2024). [RR2024-2] 교육용 인공지능 시스템의 학교 현장 적용을 위한 기초연구. 서울: 한국교육학술정보원.

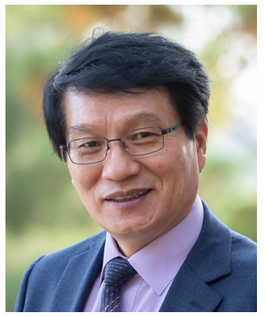
• 한국청소년정책연구원(NYPI). (2024). 청소년의 생성형 AI 이용 실태 및 리터러시 증진 방안 연구. 연구보고 24-기본-02. 세종: NYPI.

• The Economist. (2025, February 13). How AI will divide the best from the rest. Finance & Economics Section. London: The Economist Group.

• 동아일보(동아시론) (2025, July 25). 불평등 부르는 'AI 디바이드', 공교육이 해결의 열쇠.

AI 대전환 시대의 교육 패러다임의 전환 필요성과 인문학적 인재상의 확립

The necessity of changing the educational paradigm and the establishment of humanistic talent in the era of the AI transformation



김원중
단국대학교 교수

Hyeoncheol Kim
Professor, Korea University

Abstract

AI로 전환되는 이 시대에 우리는 인문학을 통해 과학 기술의 방향이 인간 중심으로 전환하여 지속 가능하고 조화로운 미래로 나아갈 수 방법을 고민해야 하고 그 해결책을 모색해 보아야 한다. 인간 중심적 가치제고와 통찰력이 더욱 요구되는 이유는 인문학이야말로 소통, 창의성, 공감 능력 등 인간 고유의 감성을 발달시키며, 특히 사회적 문제로 대두될 개연성이 큰 AI의 윤리 문제에 대한 해법도 제공할 수 있기 때문이다.
인문학적 인재상의 목표는 기본적으로 조화와 상생을 위한 기반 마련이요, 범세계적으로 첨예한 갈등의 지속 양상 속에서 소통과 화해롭고 균형잡힌 인성을 갖춘 인재를 양성하기 위한 기본 전제가 될 수 있다. 결론적으로 전인적 창조력과 상상력있는 인재로 양성하기 위해서는 인문학적 인재상이 필요하며, AI 대전환 시대에 있어서 심화되는 제반 문제를 해결하기 위한 최소한의 장치이기도 하다.

Abstract

In this era of AI transformation, we must contemplate through the humanities how to redirect scientific and technological development toward human-centered approaches, thereby advancing toward a sustainable and harmonious future, and explore viable solutions. The imperative for enhanced human-centered values and insight stems from the humanities' capacity to cultivate quintessentially human sensibilities such as communication, creativity, and empathy, while particularly providing ethical frameworks to address AI-related dilemmas that possess significant potential to emerge as critical societal concerns.
The objective of humanities-based human development fundamentally establishes foundations for harmony and symbiosis, serving as an essential prerequisite for cultivating individuals equipped with communicative competence, conciliatory dispositions, and balanced character amid the persistent patterns of acute global conflicts. In conclusion, humanities-oriented human development is indispensable for nurturing individuals with holistic creativity and imagination, constituting a minimal yet crucial mechanism for addressing the intensifying complexities inherent in the era of AI transformation.

1. 왜 인문학적 인재상이 필요한가

AI로 전환되는 이 시대에 인문학의 가치가 더 높아질 수 있는가? 아니면 이런 과학 기술의 발전이 인문학의 위기를 심화시킬 것인가? 적어도 고등교육기관의 최상위권 학생들의 압도적 대다수가 의과대학으로 몰리는 대한민국의 참담한 인문학 입지 여건과 진학 현실을 고려해볼 때 현재 직면한 문제는 후자의 경우에 속한다는 데 거의 동의할 것이다. 그런데 우리의 이런 현실과는 현저히 다른 움직임이 있었으니 오픈 커리큘럼을 통해 인문학의 기반하에 사회 과학과 자연 과학 등을 함께 공부하여 융합적 사고를 키우는 것으로 유명한 미국 최고의 인문대학 중의 하나가 바로 애머스트 대학이다. 그곳의 엘리엇 총장은 "많은 문제를 해결하는 데 AI가 도움은 되겠지만 어떤 질문을 던질지는 인간이 결정해야 한다"고 믿음 때문인지 그 대학은 6명의 노벨상 수상자를 배출했다고 한다.

이 말에 동의하든 않든 우리는 인문학을 통해 과학 기술의 방향이 인간 중심으로 전환하여 지속 가능하고 조화로운 미래로 나아갈 수 방법을 고민해야 하고 그 해결책을 모색해 보아야 한다. 인문학이야말로 인간의 존엄성에 대한 복잡한 윤리적 딜레마를 극복하고 해소하는 역할을 한다는 이유일 듯한데, 시대와 공간을 불문하고 인문학은 인간의 가치에 대해 본질적이고 근본적이며 다양한 맥락에서 생각할 여지를 제공하는 학문이기에 가능한 것이다. 사실상 교육 패러다임의 변화—형식적 지식(explicit knowledge)이냐 암묵적 지식(tacit knowledge)이냐 하는 문제가 계속 대두되고 있으며, 생성형 인공지능과 초연결(hyper-connectivity) 시대(*빅데이터, 인공지능, IoT 등 획기적인 지식정보 혁명)의 시기의 초지능과 초학제적 특성을 보이는 현실임을 감안하면 챗GPT의 출시(2022년 11월)로 인한 교육 패러다임은 변화되고 있고, 수월성의 새로운 모델을 요구하는데 유능하고 유덕한 휴머니즘이 있는 인재상이 요구되는 현실임에도 불구하고 기계가 인간을 지배하는 시대가 오고 있다. 그렇다면 인간은 기계가 하는 일을 하지 않고 하지 못하는 일을 해야 하는데, 과연 그 해답이 어디에 있는가 하는 점이 필요하다는 것이다.

늘 우리는 인구절벽과 기후환경 및 교육 환경의 변화는 문명사적 대전환기라고 불리는 격변의 시기가 당도했음 직시하게 된다. 여기에 발맞춰 과연 어떤 교육 패러다임의 모색이 진행되고 있으며, 이런 시점에서 왜 인문학적 인재상이 필요한지 점검할 필요가 있다. 디지털에 지배되어 타인과의 가시적 소통이 거의 무력화되는 현실에서 타인의 문화와 역사 및 언어의 복잡성을 이해할 수 있는 인재상이 무엇이며 왜 인문학적 인재상의 확립이 왜 요구되는지 알아볼 필요가 있다.

이미 고인이 되었지만, "인간의 감동은 인문학을 기술로 구현해야"한다는 애플 창시자 스티브 잡스가 꽤 오래전에 던진 명제가 여전히 유효하다고 한다면, 그 이유는 물질 만능 시대의 오늘을 사는 경직된 제도하에서 기계적이고 표준화된 교육과정의 근본적 변화를 맞이함에 있어서 인문학의 위상이 중요하다는 인식이니, AI가 중심인 현 상황에서도 인문학과 조화의 가능성은 유효한 명제를 제시하고 있다. 물론 절망적인 보고도 있다. 세계경제포럼의 <직업의 미래보고서 2023>에 의하면, 세계 45개국 803개 기업 대상 조사 결과, 향후 5년간 AI 기술의 눈부신 발달과 사용 급증으로 기존 일자리 25%가 변화하고 2,600만 개의 일자리가 사라질 위기라고 전망하고 있다.

주지하듯 융합 교육이 대세인 지금의 상황에 걸맞게 "기술통합/학제간 접근방식/혼합학습/글로벌 교육표준/데이터 기반 의사결정/협동학습" 등 6가지가 융합의 핵심이라고 챗GPT도 설명하고 있는데, 감각적 지식이 논리적 사유보다 중요하다는 의미가 여기엔 담겨 있다. 역설적으로 문제해결 능력의 실종이 가속화될

수록 통합적이고 융합적 사고 능력은 줄어들고, 모든 지식은 편협하고 협소한 데로 향할 개연성과 위험성이 도사리고 있다. 말하자면, AI 딥페이크와 관련한 모든 것이 위기(학습 아바타, 데이터 아바타)인 현실에서 과연 인간의 실종가능성은 현저해지고 심지어 존재조차 사라지게 된다면, 과연 우리는 어떤 것이 나은 선택지가 되어야 하는지 고민해야 한다.

따라서 인간 중심적 가치와 통찰이 더욱 요구되는 이유는 AI가 주축으로 하는 데이터 처리와 효율 중심의 사고라면 인문학은 소통, 창의성, 공감 능력 등 인간 고유의 감성을 발달시키며 특히 사회적 문제로 대두될 개연성이 큰 AI의 윤리 문제에 대한 해법을 제공할 수 있기 때문이다.

즉 시대에 따라 다양하게 변화하는 문화, 인간 내면의 갈등과 고민을 탐구함으로써 이해와 공감을 키우는 인문학의 본질적 가치야말로 성공을 위한 수단적 · 기술적 차원이 아닌 인간 중심의 교육의 근본적인 목표를 지향한다는 점을 말이다. 즉, 인문학은 그 주된 기능이 지식 중심의 언어가 아닌 사람의 가치에 주목하고자 하는 것이 일차적인 목적이며, 이는 인간의 윤리적 판단과 사회적 책임을 확고히 하기 위해 인문학이 지향하고 있는 윤리적 기준과 사회적 책임에 대한 논의를 촉진하는 역할을 하기 때문에 아무리 시대가 급변해도 인문학에 기반을 둔 인재가 필요하다는 논리가 설득력을 얻지 않을까?

2. 인문학적 인재상이 필요하고 긴요하다

교육의 패러다임 변화로 인해 다양한 가치와 생각이 충돌하고 새로운 지식과 정보가 계속 생성되는 지금의 상황에서 사회 통합과 공동체 발전을 위해 인문학적 인재상의 필요성은 더욱 중요해지고 있으니, 인문학 학습을 통해 구성원간에 조화와 협력을 추구하게 되어 사회의 갈등 지수를 완화하는 데에도 기여할 수 있다.

그러므로 우리가 일관되고 지속되게 추진해 온 교육의 순서인 '지덕체知德體'가 아니라 '덕체지德體知'로 교육 패러다임이 다시금 전환할 필요가 있다는 논의도 설득력이 있다. 이는 국가교육회의가 실시한 2022 개정 교육과정 관련 국민 참여 설문 조사에 따르면 초·중·고등학교에서 가장 강화되어야 할 교육으로 인성교육(36.3%)이 가장 중요한 항목으로 선정된 것과 맞물리며,국민참여위원회('23.8%), 미래 사회에 필요한 교육으로 인성교육을 가장 많이 응답(18%)한 데서도 입증 여력이 있다는 의미다. 덧붙여 주지하듯 국제학업성취도평가에서 한국 학생들이 높은 성취를 보이고 있음에도 불구하고, 각종 데이터 자료를 통해 입증하고자 하듯 삶의 만족도는 OECD 평균보다 현저히 낮다는 점에 동의하는 학자들의 견해도 설득력을 얻는다.

요컨대, 불균형의 원인을 인지적 역량에만 집중한 교육 구조에서 찾고 있으며 이제는 비인지적 역량—자기조절, 회복 탄력성, 사회성, 공감, 인성과 시민성-을 강화해야 한다는 논의가 설득력을 얻고 있다.[1] 이런 논의가 설득력을 얻기 위해서는 미래 사회에 필요한 인재를 양성하기 위한 기초를 다지는 방안으로 인문학적 지혜가 절실하다는 인식하에 동서양의 인문학의 지혜 습득을 기본으로 한 인문학적 인재상이 미래 창의성의 기반 마련될 수 있다는 것이며, 그러기에 인문학은 물론 과학과 예술이 융합된 교육 기회 마련될 수 있다는 점이다.

지금도 그렇고 다가오는 수년 내에 예측할 수 없는 문제들이 많이 발생할 것임을 의심하는 이는 아무도 없고 이래서 창의적 사고와 문제해결 능력을 길러야 한다고 외친다. 그렇다면 창의적 사고와 문제해결 능력을 키우는 교육이 필요하다면 문제를 발견하고 분석하며, 새로운 아이디어를 생각하고 구현하는 능력이 무엇이

1) 이런 논의가 설득력을 얻기 위해서는 미래 사회에 필요한 인재를 양성하기 위한 기초를 다지는 방안으로 인문학적 지혜가 절실하다는 인식하에 동서양의 인문학의 지혜 습득을 기본으로 한 인문학적 인재상이 미래 창의성의 기반 마련될 수 있다는 것이며, 그러기에 인문학은 물론 과학과 예술이 융합된 교육 기회 마련될 수 있다는 점이다.

냐는 것인가 하는 문제로 집약된다.

현재 한국의 경우에도 학문간의 경계를 허무는 노력이 가속화되어 무학과, 무전공, 무학년 등의 개념이 도입되어 거의 모든 대학에서 일정 비율을 할당하여 학생들에게 기회를 제공하고 있는데, 이러한 시도의 결과에 대해서는 수많은 비판과 논란이 지속되고 있고 그 성과는 물론 최소 4년 후나 6년 이후에 공과를 평가할 수 있을 만큼 이런 시도는 일관된 방향성 없이 갑자기 도입되어 그 성과를 보장하기 힘들 것이다.

3. 휴머니즘의 인재상은 인문학 기반 교육에 기반

비판적 사고와 융합적 역량을 키우기 위해서는 인문학이 필요한 것은 당연하며, 창의적이고 직관적인 사고방식은 문제해결 능력을 향상시키고 호기심에 바탕을 둔 책임감이 필요한데, 이는 사회적 책무이기도 하며 21세기 현 교육 패러다임의 모델이기도 하다. UNESCO 및 OECD 미래교육 전망(2022)에서 "예측 불가능한 미래 사회는 인문학적 소양을 갖춘 시민적 책임의식과 탐구적 자율인이 필요하며 이는 협동과 협력 연대 포용 등"을 언급했듯이 공유 지식과 디지털 지식의 이중성을 강조하는 교육 방향과도 관련해서 함께 생각해 보아야 한다. 따라서 미래의 변혁적 역량을 확대하기 위한 교육 패러다임의 변화에 중심 역할을 하는 것이 어떤 인재상을 정립하느냐 하는 문제인데 바로 인문학적 소양을 갖춘 휴머니즘 인재상일 것이다.

우선, 인문학적, 사회학적, 과학기술적 관점 등 다양한 시각에서 제기된 교육 방향을 디지털 만능 사회인 작금의 상황에서 교육의 기본 패러다임의 급변을 인정하고 글로벌 스탠다드에 걸맞는 시야의 확보가 필요하다는 점을 인지할 필요가 있으며, 체계적인 인문학 교육을 통해 원만한 인성과 시민성을 함양할 수 있도록 지속적인 교육 지원과 평가 피드백이 필요하다.[2)]

왜냐하면 AI의 기본적인 고질병으로 대두될 윤리적 문제에 대한 대비수단으로서 인문학적 성찰이 요구되는 현실에서 출발하여 디지털 만능의 보완책이 무엇인지 고민해야 하고 그 해결책 모색을 위해 노력해야 한다. 인문학을 통해 인성 역량도 함양되고 공감 능력과 공동체에 대한 관심과 배려 및 갈등의 조정과 해결 능력의 함양은 인문학 교육을 통해 일정 부분 가능하다고 보기 때문이다.

이렇게 강조하는 이유는 인문학적 인재상의 목표는 기본적으로 조화와 상생을 위한 기반마련이요, 현재 우리나라 전반에 걸쳐 있는 첨예한 갈등의 지속 양상 속에서 소통과 화해롭고 균형잡힌 인성을 갖춘 인재를 양성하기 위한 기본 전제이기도 하기 때문이다. 요컨대, 휴머니즘과 포용성 및 다양성임을 알고 이를 달성하기 위해 우리는 인간과 사회에 대한 성찰적 대응이 바로 인문학적 인재상을 구축할 필요가 있는 것이다. 이는 다시 말해 디지털과 인문학의 역량을 갖춘 융복합 능력을 갖춘 인재상이어야 한다. 한쪽을 도외시하고 다른 편만 강조하면 제대로 된 인재가 나올 수 없다는 위기의식에 다들 동의하면서도 여전히 인문학적 인재상 양성에는 반대표를 던진다.

과연 우리의 행복지수는 어떠한가 거론할 필요조차 없지 않은가? 주지하듯이 OECD 국가중 최상위권을 차지 하고 있는 청소년의 불행지수나 자살율 등이 입증하고 있지 않은가 말이다. 모든 국민은 자아실현과 타고난 능력을 발휘할 수 있는 교육을 받을 권리가 있으며, 이는 시대와 현실에 능동적으로 대응할 수 있는 유연한 인재 발현의 갈망이기도 한데, 그 유연한 인재상의 확립에 인문학 교육이 절실하게 요구되고 있다는 데 동의하면서도 그것을 외면하는 이유가 바로 인문학을 통해서는 먹고사는 문제에 대한 확신이 없다는 인식이 유독 우리에겐 강하다는 인식이 거의 사회 전반에 굳건하게 자리 잡고 있다는 데 연유한다. 인성을 갖춘 최소한의 교육이 인문학적 인재 양성에 있고, 자기 주도성, 협업, 의사소통, 정보 해결 능력, 공감 능력 등을 고양시킬 수 있는 뿌리도 인문학에 있는데 이것에 대한 고민이나 성찰의 기회조차 없고 그럴 여유도 없다는 말이다.

물론 그렇다고 해서 수학과 과학 등 인지능력(기본학력)이 있어야만 창의성, 문제 해결력, 협업 능력도 나온다는 것을 부정하는 것이 아니며, AI 시대가 온다고 해서 인지능력의 중요성이 줄어들지는 않고 오히려 인지능력은 더욱 강화될 수 밖에 없다. 인지적 역량 즉 학력도 절대 소홀히 할 수 없고 그것을 외면해서도 안된다. 학력과 인성의 확장을 위한 인문학 교육 등을 함께 공존하게 하면서 교육하는 방법을 모색해 봐야 하는 것이며, 그런 모색은 적극적으로 모색하자는 것이니, 지금 한국에서도 아주 일부 시도(태재대학이 그 한 예다)들이 있기는 하다.

복합적 지식과 능력에 대한 요구로서 복잡한 문제를 해결하기 위해서는 지식의 위계성보다 지식의 총괄성이 중요하며, 실제 생활에서 '무엇인가를 행할 수 있는 능력'은 당연히 요구되는 것도 자명하다. 이것은 암묵적 지식(tacit knowledge)으로 교육 내용의 융합으로 갖게 되는 통합적 능력이기 때문에 휴머니즘 교육이 중요하다는 측면에서 상호 연결되고 효과적 학습 환경을 만들기 위한 '컨버전스' 교육 방식 중심에 휴머니즘이 있어야 한다는 것이다. 왜냐하면 사회적 변혁은 인문학을 통해 학제가 발달해가는 데 바로 포용과 협력, 연대 및 공존의식 확장 가능성이 탁월한 휴머니즘 정신에 배어 있기 때문이다.

기본에 충실한 제대로 된 사람을 만드는 인성 교육이 제대로 이루어지기 위해서는 인문학적 인재 양성에 대한 사회 구성원의 합의가 필요하며, 학부모와 학생도 소위 경제적으로 보장된 특수한 과를 지망하기보다는 나아가 근본적인 국가의 장래와 미래를 생각하고 학과를 결정하는 안목을 길러야 한다.

4. 결론은 인문학이다

비판적 사고력, 창의력, 표현력, 소통, 협력, 사회적 책임감 등 인문학 교육과 관련한 많은 용어들은 인문학의 기반 교육을 통해 키워지는 핵심 역량을 반영하고 있다. 인문학 교육은 단순히 지식 전달의 수준을 넘어 이러한 역량을 기반으로 교육 패러다임의 전환이 필요하다. 그런데 오늘날 교육 현장에서 정답이 다양하게 존재한다는 것을 배우면서 우리는 사고의 폭을 넓히고 문학과 역사 및 철학을 통해 창의력과 자기 표현 능력을 향상시키게 된다. 그리고 토론식 수업을 통해 타인과의 소통능력을 배우는데 이 과정에서 사회적 공존의 중요성을 체험하게 된다.

건전한 경쟁과 협력을 경험하는 교육은 중요하며 인간다움과 정의와 공정 등 보편적 가치에 대한 성찰이 필요하다. 테크노와의 공존을 위해 인문학을 통해 인간 문명에 대한 압도적인 힘을 비축할 수 있으며, 고도의 '공감(empathy)' 능력과 '감수성(sensitivity)'을 갖춘 인재는 초지능의 주도적 활용 능력을 갖춘 유연함과 회복 탄력성을 갖춘 통섭적 인재로 성장하며 자유 민주시민으로서 사회적 책무와 윤리적 판단 능력을 갖추게 될 것이다. 거의 지식 전달의 차원에서 이루어진 교육에서 AI로 급변하는 패러다임의 전환 속에서 길잃은 인문학적 인재상에 대한 성찰은 미래 교육이 인지 역량과 비인지 역량의 확보가 균형이 요구된다는 논리에서 출발하며, 그런 균형이 완전히 무너지고 나면 우리의 행복은 더 이상 회복 불능할지도 모르기 때문이다.

2) 물론 이런 시도는 공교육이 거의 붕괴된 실제 공교육 현장에서 수용되지 않을 개연성이 있어 거론하는 것 자체가 무의미하다.

특히 사회의 보편적 질서 규범에 공감하고 실천하기 위한 인문학 교육이 필요함에도 불구하고 삶의 질과 다양성을 추구하고 4차 산업, AI 시대의 선도국가의 핵심 DNA가 어디에 있는지 충분히 고민하고 성찰해야 한다. 인문학을 통해서 말이다. 말로만 하는 것이 아니다. 날로 심화되어 가는 인간성 상실을 벗어나 공감과 배려의 따스함, 역동성이 있고 도전과 선도적 인재상 제시를 위해 방향성과 협조성을 이끌어내는 힘이 어디에서 나오는가 하는 것은 자명하지 않은가? 한걸음 더 나아가 급격한 기후변화, 에너지 고갈, 전쟁 위기, 양극화, 팬데믹 등의 현 상황에서의 유연한 대응 전략도 인문학에 있다고 본다. 결국 생존과 경쟁을 넘어 행복지수를 높이기 위한 시대적 요구에 부응할 수 있으며, 인류 공동체의 지속 가능한 삶에 기여하는 근본적인 문제를 고민할 수 있는 계기를 마련할 수 있는 교두보를 확보해야 한다는 의미다.

간단하게 살펴본 것처럼, AI와 같은 디지털 기술의 발전으로 인해 겉으로 보기에는 사회적 네트워크가 확대되어 글로벌 시대인 듯 하지만 우리의 행복지수는 나날이 감소하고 있지 않은가? 왜 그런가 이유를 따져보면 상대방을 존중하고 배려하면서 소통하는 인성의 부족이 자리 잡고 있지 않는가? 거의 최상위권을 차지하는 저출생국가라는 오명과 고령화, 심지어 다른 어느 나라보다도 지역소멸 현상이 만연하고 수도권 중심이 심화되는 현상이 지속되고, 환경 · 기후변화, 혐오 · 차별의식의 확산 등 사회 공동의 문제 극복을 위해서 가치관의 전환 교육이 필요하지 않은가? 타인, 공동체, 자연과 더불어 살아가는 데 필요한 가치와 덕목의 내면화가 필요하기에 인문학 교육은 올바른 인재상의 확립에 최우선적으로 선행되어야 할 필요조건이라고 할 수 있다.

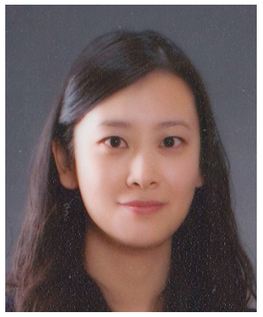
결론적으로 말해서 파편적 지식과 암기력 중심 교육이라고 비판받는 교육의 한 단면을 전인적 창조·상상력으로 전환하기 위해서는 인문학적 인재상이 필요하며, AI 대전환 시대에 있어서 심화되어 각종 폐단을 예방하기 위해서는 보다 적극적으로 인문학적 인재상을 도입하는 것은 그 어느 때보다 절박하고 긴요한 시대적 책무라고 본다.

분과회의 세션 04
Parallel Session 04

AI와 번역
Education in the Age of AI

AI 기반 한문 번역의 현황과 전망

Current Status and Prospects of AI-Based Classical Chinese Translation



임영길
성균관대학교 교수

Younggil Yim
Professor, Sungkyunkwan University

※ 해당 강연자의 초록 및 논문은 현장 발표로 대체합니다.

THE 8th WORLD HUMANITIES FORUM
제8회 세계인문학포럼

분과회의 세션 4
AI와 번역

Parallel Session 4
AI and Translation

인간-AI 협력이 대형 언어 모델의 번역 성능에 어떤 영향을 미치는가? — 사전 연구 —

How Does Human-AI Collaboration Affect the Translation Performance of Large Language Models? A Preliminary Study

커 휴
맬버른대학교 교수

Ke Hu
Professor, The University of Melbourne

Abstract

This preliminary study investigates how different human–AI collaboration strategies affect the translation performance of large language models (LLMs). Chinese–English translations of a legal text and a literary text were generated by ChatGPT-4o in three workflows with different levels of human-AI collaboration: (1) a simple prompt with minimal task information, (2) a detailed prompt with context-specific human instructions, and (3) a detailed prompt with context-specific human instructions and a bilingual glossary. For legal translation, results show that, compared with the simple prompt, the more detail prompts substantially reduced both the number and severity of errors, especially in Terminology and Accuracy. For literary translation, error analysis suggests that the context-rich prompt failed to lower the overall error count, whereas a kudos-based evaluation revealed that context-specific instructions elicited a notable number of creative, non-literal translation solutions that contributed to better readability. Overall, these findings show that human-provided contextual information and bilingual glossaries are effective in increasing the quality of LLM-generated translations for the legal and literary texts examined.

Introduction

Since the public release of OpenAI's large language model (LLM) ChatGPT in November 2022, the rapid rise of artificial intelligence has been reshaping translation practices. When applied to translation, one of the key advantages of LLMs over earlier machine translation (MT) systems is their ability to customise translation output in response to user input. For a given source text, traditional MT tools such as Google Translate and DeepL typically generate only a single fixed translation, or at best a limited set of alternatives. By contrast, a human user can potentially use an LLM to generate an unlimited number of alternative translations by giving the AI model different instructions (i.e., prompts).

In the broader field of human-AI collaboration, an effective prompt has been found to be crucial for improving LLMs' performance in various tasks (Ziegler & Berryman, 2023; Fulford & Ng, 2023). In non-translation tasks such as programming and text summarisation, ambiguous prompts have been found to generate irrelevant or misleading responses (Jiang et al., 2022), whereas structured, context-rich instructions can reduce bias and error in LLM outputs (Dong et al., 2023; White et al., 2023).

To date, only a limited body of empirical research has initially explored how different prompt-construction strategies affect the quality of LLM-generated translation. Collectively, these studies suggest that LLMs' performance on translation tasks tends to improve when the user prompt includes exemplary translation examples (Jiao et al., 2024; Pourkamali & Sharifi, 2024; Zhang et al., 2023a; Zhang et al., 2023b) and task-related information about the source text, the target audience, or stylistic requirements (Gao et al., 2023; Jiao et al., 2024; Yamada, 2023).

Despite these initial findings, previous studies have only sporadically investigated a few isolated types of task information, without systematically incorporating all relevant aspects of a translation task into the prompt. Building on these works, Hu (in press) is the first study to propose a comprehensive framework for designing effective prompts for translation tasks. Focusing on Chinese–English legal translation, the study compared four workflows of human–AI collaboration that ranged from low to high levels of human-AI collaboration. The findings showed that the quality of LLM-generated legal translations improved when the large language model was provided with a prompt containing context-rich instructions and a human-verified bilingual glossary.

In Hu (in press), a "STAR-STAR" framework is proposed as the first comprehensive model for designing effective prompts for translation tasks. Building on previous works (e.g., Nord, 2008, 2018) on translation briefs for human translators, the STAR–STAR framework sets out eight task-specific dimensions that should be considered when instructing a large language model for translation:

- Source text information,
- Target audience,
- Aim of translation,
- Requirements specific to the context,
- Style of the target text,
- Terminology,
- Approach to translation, and
- Response format.

While Hu (in press) examined the effectiveness of the STAR-STAR framework specifically in the domain of legal translation, the framework's broader applicability remains to be tested. This preliminary study forms the first step in a larger research project that aims to evaluate the effectiveness of the STAR-STAR prompting framework across a wider range of translation domains. In this paper, I compare Chinese–English translations of legal, healthcare, and literary texts produced by ChatGPT-4o using a simple prompt versus a context-rich STAR–STAR prompt.

Methodology

Research aim

The primary aim of the broader research project to which this paper contributes is to explore a comprehensive framework of human–AI collaboration across multiple translation domains. Specifically, this preliminary study examines whether a context-rich prompt constructed based on the STAR–STAR framework (Hu, in press) improves the quality of AI-generated translations of legal, healthcare, and literary texts, compared with outputs generated from a simple prompt containing no contextual information.

Data Selection

To provide an initial test of the cross-domain applicability of the STAR–STAR framework, two distinct translation domains were selected: legal and literary. These domains were chosen because they pose markedly different translation challenges. Legal texts are characterised by specialised terminology, complex syntax, and conceptual asymmetries across different legal systems (Cao, 2007). By contrast, literary texts present a high demand for creativity, stylistic aesthetics, and cultural nuance. Specifically, translation examples in this study were drawn from two representative Chinese source texts:

- ST-A: First two chapters of The Civil Code of the People's Republic of China.
- ST-B: First two chapters of the web novel魔道祖师 (Grandmaster of Demonic Cultivation).

Translation workflows

All translations were generated using OpenAI's ChatGPT-4o model. Three translation workflows were adopted for each source text:

Low human input:

In this workflow, each source text was translated by ChatGPT-4o using a simple prompt carrying minimal information on the translation task: "Translate the following text into English". The resulting translations are henceforth named TT-A1 and TT-B1.

Medium human input:

Immediately after the generation of TT-A1 and TT-B1, ChatGPT was instructed to re-translate each text using context-rich prompts based on the STAR–STAR framework. For ST-A, the context-rich prompt was designed as follows (Hu, in press, p.14):

Re-translate the text according to the instructions below:
###
Instructions:
#Source text information:
The source text is the Civil Code of the People's Republic of China (中华人民共和国民法典), which provides a comprehensive legal framework for civil affairs in China.
#Target audience:
The translation is intended for English-speaking legal practitioners, legal scholars, and non-expert stakeholders.
#Aim of translation:
The translation must faithfully represent the legal meaning of the source text to provide the English-speaking audience with a clear and accurate understanding of China's civil law.
#Requirements specific to the context:
Do not introduce any errors or ambiguities that could lead to a misunderstanding of the legislative intent of the source text.
Accurately convey the logical connections in each legal article.
Note that the source text is under the continental law system. For a Chinese term that does not have an English equivalent due to the gap between continental and common law systems, explain the term in English instead of using a misleading concept from common law.
#Style:
Use a formal, impersonal writing style.
Adhere to the stylistic conventions of civil law statutes.
#Terminology:
Use widely accepted English terminology in civil law.
Use the same term for the same legal concept to prevent misunderstanding.
#Approach to translation:
If a literal translation of any term or clause results in a misunderstanding of the original legal meaning, make appropriate lexical and syntactic adjustments to ensure clarity and accuracy.
#Response format:
Translate all sections of the source text.
Preserve the numbering, headings, and subsections as they appear in the source text.
###

For ST-B, the STAR-STAR prompt was:

Re-translate the text according to the instructions below:
###
Instructions:
#Source text information:
The source text is the Chinese Xianxia web novel《魔道祖师》, set in an imagined ancient cultivation world with complex character relationships and rich cultural references.
#Target audience:
General English-speaking readers with limited knowledge of the Chinese culture or the Xianxia genre.
#Aim of translation:
Enable general readers to follow the story, appreciate its cultural and emotional appeal, and enjoy an engaging reading experience.
#Requirements specific to the context:
Do not introduce any errors or ambiguities that could distort the plot, character relationships, and themes of the source text.
Do not introduce any omission or addition, unless it is required for clarity, readability, or other reasons essential to a successful translation.
#Style:
The translation should convey both the sense of antiquity in its classical elements and the accessibility of its colloquial voice.
The translation should recreate the author's fast-paced, vivid, emotionally rich, and humorous style with lively, satirical dialogue.
#Terminology:
Refer to the Xianxia genre and Chinese cultural context, ensuring that character relationships, names, and historical elements are translated appropriately.
For terms unique to this novel, apply flexible strategies balancing meaning, cultural preservation, aesthetics, and readability.

Pay attention to the nuanced differences between similar terms in the source text.
Ensure terminology consistency in the target text.
#Approach to translation:
Use literal translation as the default. Adjust minimally only if it would cause misunderstanding, ensuring clarity, accuracy, and fluency.
Use an explicitation strategy for any elements difficult to understand.
#Response format:
Translate all sections of the source text.
Keep the original heading structure, but translate the headings and subheadings.
Apply boldface formatting to all headings and subheadings.
###

The translations resulting from this workflow are henceforth named TT-A2 and TT-B2.

High human input:
Following the generation of TT-A2 and TT-B2, each source text was further re-translated by ChatGPT using the same STAR-STAR prompts from the previous workflow, supplemented with a human-verified bilingual glossary that comprised terms relevant to each source text. The translations resulting from this workflow are henceforth named TT-A3 and TT-B3.

Evaluation methods
In this preliminary study, qualitative analysis was employed to gain fine-grained insights into the quality differences between translations generated from the three workflows. For both ST-A and ST-B, translation outputs were examined using the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2024), a widely adopted model for assessing translation quality. Errors were classified according to the MQM typology across seven main categories: terminology, accuracy, style, linguistic conventions, locale conventions, audience appropriateness, and design/markup. Each error was additionally assigned to two severity levels: minor and critical.

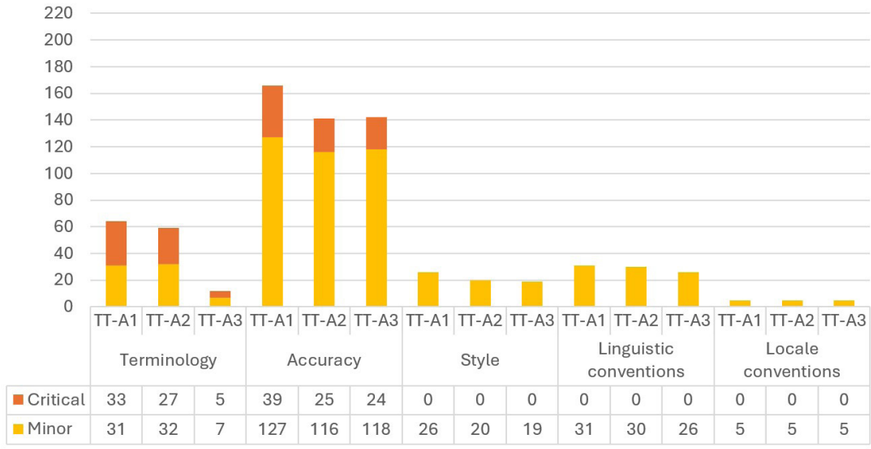
While analysing the LLM-generated literary translations for ST-B, it was found that the context-rich prompts not only affected the number of translation errors, but also introduced rewardable creative translation adjustments that contributed to the readability of the translations. Given that the error analysis falls short of capturing the quality increase brought about by these creative translation adjustments, a Kudos-based evaluation approach was adopted for ST-B to complement the penalty-based error analysis. Instead of penalising errors, the Kudos-based evaluation identified creative solutions that improved readability, particularly non-literal interventions such as addition, omission, and adaptation.

In the case of ST-B, four main types of rewardable strategies emerged from the Kudos-based evaluation: explanatory addition, descriptive addition, rhetorical addition, and coherence addition. These categories capture the ways in which LLM outputs introduced creative adjustments that went beyond literal transfer to enrich the narrative and enhance readability. By recognising these rewardable translation interventions alongside translation errors, the qualitative analysis yields a more nuanced and balanced evaluation of LLM-generated literary translations.

Results and Discussion

Results for legal translation

The MQM-based error analysis reveals clear differences across the three workflows of legal translation (Figure 1). Produced with a simple prompt, the low-human-input version TT-A1 contained the highest number of critical errors, especially in Terminology and Accuracy. In this version, a large number of legal terms were translated incorrectly or inconsistently, and the rendering of complex sentences often distorted the original meaning.



	TT-A1	TT-A2	TT-A3	TT-A1	TT-A2	TT-A3	TT-A1	TT-A2	TT-A3	TT-A1	TT-A2	TT-A3	TT-A1	TT-A2	TT-A3
	Terminology			Accuracy			Style			Linguistic conventions			Locale conventions		
Critical	33	27	5	39	25	24	0	0	0	0	0	0	0	0	0
Minor	31	32	7	127	116	118	26	20	19	31	30	26	5	5	5

Figure 1 MQM-based error analysis for three legal translation workflows

By contrast, the use of a STAR–STAR prompt (TT-A2) reduced the overall number of errors, especially critical ones in the Accuracy and Terminology categories. This indicates a context-rich prompt with detailed human-provided instructions can effectively improve the overall quality of the legal translations.

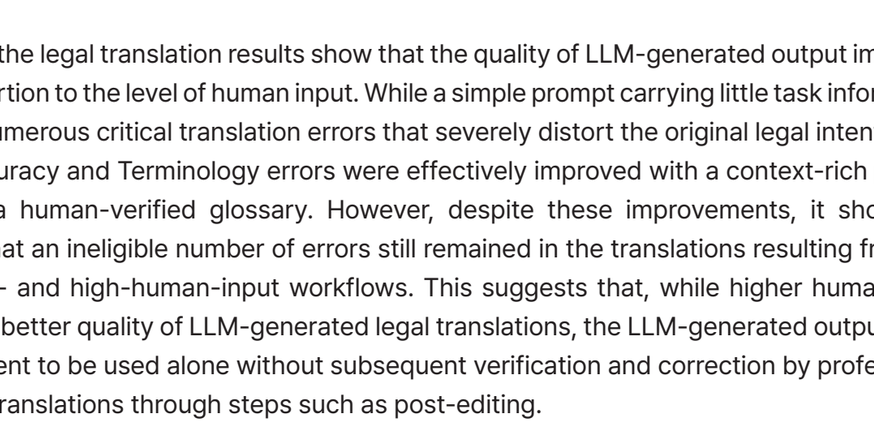
The greatest improvement was observed in the high-human-input version (TT-A3), where the LLM was given both the context-rich prompt and a human-verified bilingual legal glossary. In this version, most key legal terms were rendered correctly and consistently in line with the provided glossary, resulting in substantially fewer terminology errors than in the

other workflows.

Overall, the legal translation results show that the quality of LLM-generated output improves in proportion to the level of human input. While a simple prompt carrying little task information led to numerous critical translation errors that severely distort the original legal intent, many key Accuracy and Terminology errors were effectively improved with a context-rich prompt and/or a human-verified glossary. However, despite these improvements, it should be noted that an ineligible number of errors still remained in the translations resulting from the medium- and high-human-input workflows. This suggests that, while higher human input leads to better quality of LLM-generated legal translations, the LLM-generated output is still insufficient to be used alone without subsequent verification and correction by professional human translations through steps such as post-editing.

Results for literary translation

The error analysis result for the literary text is shown in Figure 2. Contrary to the results for legal translation, when a STAR–STAR prompt was used to re-translate the literary text (TT-B2), it unexpectedly produced a larger number of Terminology and Accuracy errors, especially critical ones. Furthermore, while the inclusion of human-verified glossary led to the fewest number of Terminology errors in the high-human-input version (TT-B3), this version also registered the highest count for Accuracy errors. Overall, the total counts of critical and minor errors in TT-B2 and TT-B3 were either comparable to or even higher than those in TT-B1 across most error categories. That is, based on the error analysis results alone, it seems that a more detailed prompt failed to increase the quality of LLM-generated literary translation.



	Terminology			Accuracy			Style			Linguistic Conventions			Locale conventions		
	TT-B1	TT-B2	TT-B3	TT-B1	TT-B2	TT-B3	TT-B1	TT-B2	TT-B3	TT-B1	TT-B2	TT-B3	TT-B1	TT-B2	TT-B3
Critical	13	20	9	4	6	6	1	0	0	0	0	0	0	0	0
Minor	3	2	2	15	17	19	1	1	2	6	3	5	1	1	1

Figure 2 MQM-based error analysis for three literary translation workflows

In the analysis of the three literary translation versions, it was observed that, although the context-rich STAR–STAR prompt did not effectively reduce the total number of errors, it introduced a considerable number of creative adjustments that enhanced readability. Given this, a kudos-based analysis drawing on previous research (Guerberof and Toral, 2022) was adopted to identify the rewardable translation solutions across all three versions.

As shown in Figure 3, the kudos analysis identified four types of creative non-literal interventions that contribute positively to the readability of the translations: Explanatory addition, Descriptive addition, Rhetorical addition, and Coherence addition. Across the three translations, TT-B2 contained the highest number of kudos, followed by TT-B3. By contrast, TT-B1 lagged far behind with only minimal instances of rewardable translation solutions. These findings suggest that while the use of the STAR-STAR prompt fell short in reducing the translations errors it guided the LLM to produce a notable number of creative translation solutions that positively contributed to the readability of the literary translation.

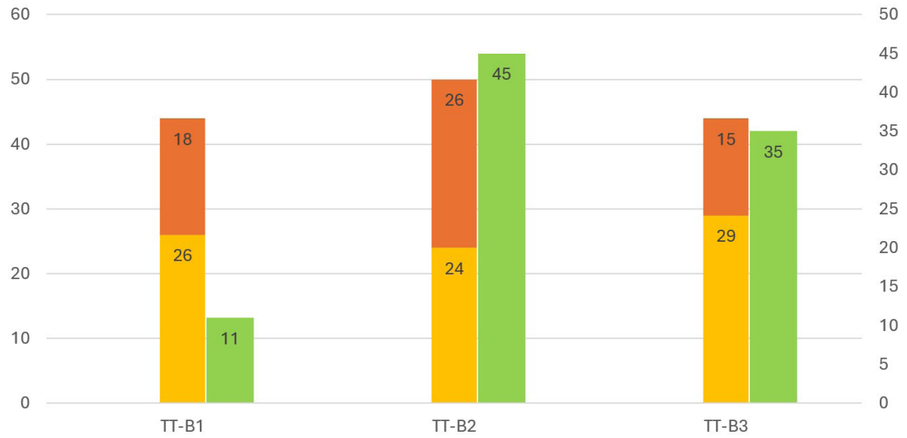


Figure 3 Kudos analysis for three literary translation workflows

Taken together, the analyses above reveal a divergence between error-based and kudos-based evaluations for the literary translation task. While error analysis indicates that a context-rich prompts unexpectedly led to more translation error, the kudos analysis shows that the more detailed prompt enhanced the LLM's translation creativity, generating a substantial increase in the rewardable creative translation solution. Given that TT-B2 and TT-B3 contained a generally similar number of translation errors but considerably more creative translation solutions than did TT-B1, the medium- and high-human-input workflows resulted in higher overall literary translation quality compared to the case of using a simple prompt.

Nevertheless, as with the case of legal translation, the outputs of TT-B2 and TT-B3 still

contain numerous errors and leave much room for improvement in stylistic aesthetics. This indicates that LLMs, even when supported by detailed prompts and glossaries, remain far insufficient for generating any directly publishable literary translation. Thus, further human post-editing is still indispensable to achieve a high-quality final translation.

Conclusion

This preliminary study seeks to examine how different levels of human–AI collaboration influence the translation performance of large language models. To this end, a comparison was made between Chinese–English legal and literary translations generated with varying levels of human input:

- Low: LLM performed the translation task with a simple user prompt with little task information;
- Medium: LLM performed the translation task with a context-rich prompt based on the STAR-STAR framework;
- High: LLM performed the translation task with context-rich prompt based on the STAR-STAR framework as well as a human-verified bilingual glossary

For legal translation, the results are relatively clear-cut: higher levels of human input led to notably fewer and less critical errors in most categories. This shows that the STAR–STAR prompt and bilingual glossary were effective in improving the overall quality of legal translations, especially in aspects such as Terminology and Accuracy.

In contrast, the analyses of the literary translations present mixed results. Error analysis alone suggested that the use of a context-rich prompt and glossary failed to reduce the total error count. However, the kudos-based evaluation highlighted the effect of human-AI collaboration on a different dimension of literary translation quality: in comparison with the simple prompt, the context-rich prompt elicited a larger number of creative, non-literal translation interventions that improved the readability of the translation. This discrepancy suggests that conventional penalty-based error analysis is insufficient to capture the qualities valued in literary translation and thus need to be complemented by kudos analysis for a more balanced evaluation for literary translation. By combining the results from the error and kudos analyses, it is found that higher human input (i.e., the use of a context-rich STAR-STAR prompt and a bilingual glossary) led to better quality of literary translations.

In summary, the findings from both the legal and literary translations consistently indicate that human-provided contextual information and glossaries are key to improving the overall quality of LLM-assisted translation. By showing that the STAR–STAR prompting framework (Hu, in press) reduced translation errors in legal texts and fostered creative solutions in literary texts, this study provides initial evidence of the effectiveness of this human-AI collaboration framework in two distinctive, yet equally challenging translation domains.

Nevertheless, it should be noted that this study is based on one translation direction (Chinese-English) and one source text from the legal and literary domains, respectively. Thus, the initial conclusions above cannot yet be generalised. Future research can extend the investigation to other high-stakes translation domains (e.g., medical translation) and more language pairs. Moreover, this study adopted qualitative analysis as the only evaluation method. In future studies, automatic translation quality metrics (e.g., the BLEU score) can be incorporated to provide complementary insights.

References

Berryman, J., & Ziegler, A. (2024). Prompt engineering for LLMs: The art and science of building large language model–based applications. O'Reilly Media.
Cao, D. (2007). Translating Law. Multilingual Matters. https://doi.org/10.21832/9781853599552
Dong, X., Zhu, Z., Wang, Z., Teleki, M., & Caverlee, J. (2023). Co2PT: Mitigating Bias in Pre-trained Language Models through Counterfactual Contrastive Prompt Tuning. In H. Bouamor, J. Pino, & K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 5859–5871). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.390
Fulford, I., & Ng, A. (n.d.). ChatGPT Prompt Engineering for Developers. Retrieved January 30, 2025, from https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/
Gao, Y., Wang, R., & Hou, F. (2024). How to Design Translation Prompts for ChatGPT: An Empirical Study. Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops, 1–7. https://doi.org/10.1145/3700410.3702123
Guerberof-Arenas, A., & Toral, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. Translation Spaces, 11(2), 184–212.
Jiang, E., Toh, E., Molina, A., Olson, K., Kayacik, C., Donsbach, A., Cai, C. J., & Terry, M. (2022). Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 1–19. https://doi.org/10.1145/3491102.3501870
Jiao, H., Peng, B., Zong, L., Zhang, X., & Li, X. (2024). Gradable ChatGPT Translation Evaluation. Procesamiento del Lenguaje Natural, 72, 73–85
Lommel, A., Gladkoff, S., Melby, A. K., Wright, S. E., Strandvik, I., Gasova, K., ... & Nenadic, G. (2024, September). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track) (pp. 75–94). Association for Machine Translation in the Americas.
Nord, C. (2008). What do We Know About the Target-Text Receiver? In A. Beeby, D. Ensinger, & M. Presas (Eds.), Investigating Translation: Selected papers from the 4th International Congress on Translation, Barcelona, 1998 (pp. 195–212). John Benjamins Publishing Company. https://doi.org/10.1075/btl.32.24nor
Nord, C. (2018). Translating as a purposeful activity: Functionalist approaches explained (2nd ed.). Routledge.
Pourkamali, N., & Sharifi, S. E. (2024). Machine Translation with Large Language Models: Prompt Engineering for Persian, English, and Russian Directions (arXiv:2401.08429). arXiv. https://doi.org/10.48550/arXiv.2401.08429
White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT (arXiv:2302.11382). arXiv. https://doi.org/10.48550/arXiv.2302.11382
Yamada, M. (2023). Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT`s Customizability. In M. Yamada & F. do Carmo (Eds.), Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track (pp. 195–204). Asia-Pacific Association for Machine Translation. https://aclanthology.org/2023.mtsummit-users.19/
Zhang, B., Haddow, B., & Birch, A. (2023). Prompting Large Language Model for Machine Translation: A Case Study. Proceedings of the 40th International Conference on Machine Learning, 41092–41110. https://proceedings.mlr.press/v202/zhang23m.html
Zhang, X., Rajabi, N., Duh, K., & Koehn, P. (2023). Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation (pp. 468–481). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.43

과거 언어, 미래 기술: 한국어 역사 자료 말뭉치와 AI 융합

Past Language, Future Technology: Integrating Korean Historical Morphological Corpora with AI

장요한
계명대학교 교수

Yohan Jang
Professor, Keimyung University

초록

본 연구는 한국어 역사 자료에 특화된 형태소 분석기 'UTagger-훈민정음'의 고도화 과정과 이를 활용한 형태소 분석 말뭉치 구축, 나아가 인공지능(AI)과의 융합 가능성을 탐색하는 것을 목적으로 한다. 기존 형태소 분석기는 현대어 어휘와 문법 규칙을 전제로 하기 때문에 중세 및 근대 한국어의 비표준 표기, 음운 변이, 특수한 어미 체계 등을 효과적으로 처리하지 못하는 한계를 지닌다. 이에 본 연구팀은 울산대학교 옥철영 교수가 개발한 UTagger를 기반으로 역사 자료 분석에 최적화된 시스템 'UTagger-훈민정음'을 고도화하여 약 60만 어절 규모의 형태소 분석 말뭉치를 구축하였다.
연구 과정에서 반복 학습 구조와 사용자 대화형 태깅 도구(TCM)를 적용함으로써 형태소 경계 인식, 품사 태깅 정확도, 미등록어 처리 등의 성능을 향상시켰다. 또한 구축된 말뭉치를 OpenAI 기반 언어모델에 적용하여 공기어 분석, 네트워크 중심성 분석, 주제별 군집 분류 및 히트맵 시각화를 시도하였다. 그 결과, 역사 자료가 디지털 인문학 연구의 핵심 데이터로 활용될 수 있으며, 교육·문화적 자원으로도 확장 가능한 가능성을 확인하였다.
본 연구는 한국어사 연구와 디지털 인문학의 새로운 연구 방법론을 제시하고, 과거 언어 자료와 미래 인공지능 기술의 접점을 마련한다는 점에서 학문적·실천적 의의를 지닌다. 다만 자료의 범위와 장르적 한계가 존재하므로 향후에는 말뭉치의 확장과 대규모 언어모델과의 연동을 통해 보다 고도화된 성과를 도출할 필요가 있다.

주제어: 한국어 역사 자료, 형태소 분석기, UTagger-훈민정음, 형태소 분석 말뭉치 구축, 디지털 인문학, 인공지능, OpenAI ChatGPT(GPT-5), 순천 김씨 묘 출토 간찰

Abstract

This study develops UTagger-Hunminjeongeum, a morphological analyzer tailored to historical Korean texts, and constructs a morphologically annotated corpus of approximately 600,000 tokens. The project further explores its applicability in conjunction with artificial intelligence (AI). Existing morphological analyzers, designed for modern Korean, face limitations in handling non-standard orthography, phonological variation, and the unique verbal endings of Middle and Early Modern Korean.

To address these challenges, we enhanced UTagger (originally developed by Professor Ok Cheol-Young, University of Ulsan) with iterative learning and an interactive tagging tool (TCM). These methods improved boundary recognition, part-of-speech tagging accuracy, and the treatment of out-of-vocabulary items. The resulting corpus was then applied to an OpenAI-based language model for collocation analysis, network centrality analysis, topic clustering, and heatmap visualization.

The findings demonstrate that historical Korean data can serve as a key resource in digital humanities, with potential applications in education and cultural heritage. This study proposes a new methodology that connects historical linguistic data with future-oriented AI technologies, thereby expanding the scope of Korean historical linguistics and digital humanities research. While the present work is limited in terms of data scope and genre, future research will expand the corpus and explore integration with large-scale language models.

Key-words: Historical Korean texts, morphological analyzer, UTagger-Hunminjeongeum, corpus construction, digital humanities, artificial intelligence, OpenAI ChatGPT (GPT-5), Suncheon Kim Clan epistolary documents
the letters excavated from the tomb of the Suncheon Kim clan

1. 서론

한국어 역사 자료는 AI(인공지능) 지능 향상의 기반이 될 뿐 아니라 역사 언어학은 물론 디지털 인문학 연구 방법에 혁신을 가져올 중요한 자원이다. 그러나 이 국어사 자료는 현대 국어와 달리 비표준 표기와 혼용 문자(한글·한자)의 양상, 시기별 음운 변이, 띄어쓰기 불안정, 다양한 고어 및 형태로 이루어져 있어 자연어처리(NLP, Natural Language Processing)을 위해서는 역사 자료에 특화된 형태소 분석 체계 구축을 위한 분석기 개발이 시급한 과제이다.[1] 이에 본 발표에서는 연구팀에서 고도화하고 있는 한국어 역사 자료 전용 형태소 분석기 'UTagger-훈민정음'의 고도화 과정과 성능을 소개하고 AI(인공지능) 기술과 결합하여 활용할 수 있는 가능성을 탐색하고자 한다.

국내외적으로 역사 언어 자료의 디지털화와 주석 작업은 점차 활발해지고 있다. 한국어의 경우, 세종계획을 비롯한 현대 한국어 말뭉치 구축은 체계적으로 진행되어 왔으나, 중세·근대 한국어 말뭉치는 제한적 규모에 머물고 있으며, 특히 형태소 수준의 주석 말뭉치 구축은 아직 초기 단계에 불과하다. 기존 형태소 분석기들은 현대어 어휘와 문법 규칙을 전제로 하기 때문에, 'ㅸ', 'ㅿ', 'ㆍ'와 같은 중세 자모, 'ㅂ계' 및 'ㅅ계' 합용 병서', 합성어 구성 방식, 중세 국어의 어미 체계 등은 제대로 처리하지 못한다. 이러한 공백을 보완하기 위해 일부 고소설 자료나 특정 장르를 대상으로 한 전용 분석기 혹은 수작업 주석 시도가 있었으나, 범용적이고 확장 가능한 체계 구축은 미진한 실정이다.[2]

현재 연구팀에서 구축 중인 'UTagger-훈민정음'은 중세 및 근대 한국어 문헌에 나타나는 고유한 음운·형태적 특성을 반영하도록 설계되어 있으며, 형태소 분석 및 주석 정보를 동시에 부착할 수 있는 태깅 기능을 갖추고 있다. 이를 통해 생산된 역사 자료 말뭉치는 디지털 인문학 연구에서 활용 가능한 확장성을 지닌다. 따라서 본 발표에서는 한국어 역사 자료에 특화된 형태소 분석기 'UTagger-훈민정음'의 고도화 과정을 학계에 공유하고, 이를 바탕으로 AI 기술과의 융합 가능성을 살펴보고자 한다. 이는 한국어학과 역사언어학 연구의 새로운 연구 방법론을 제시할 뿐만 아니라, 인문학적 텍스트를 디지털 기반에서 재해석할 수 있는 폭넓은 응용 가능성을 열어줄 것으로 기대된다.

2. 'UTagger-훈민정음'의 고도화와 역사 자료 형태소 구축 체제

2.1. 'UTagger-훈민정음'의 알고리즘과 고도화

역사 자료 형태소 분석기는 어절 및 형태소 경계 인식률과 품사 태깅의 정확도, 생소한 단어 처리(Out-Of-Vocabulary), 한자 매핑 정확도(한자-음 대응) 등의 능력을 향상하고 시기나 문헌 유형 간 도메인 이동에 견고한, 그리고 AI 응용에 쉽게 연결할 수 있는 프로그램을 개발하는 것이 중요한 일이다. 따라서 이러한 완성도 높은 역사 자료 형태소 분석기 개발을 위해서는 체계적인 텍스트 자료 구축과 정교한 알고리즘 개발이

1) 한국어 역사 자료는 단순한 언어학적 연구를 넘어, 특정 시기의 사회사·문화사 이해를 위한 핵심 텍스트이다. 불경 언해, 문학 작품 언해, 『소학언해』를 비롯한 사서삼경 언해, 한자·외국어 학습용 언간(諺簡), 왕실 윤음(綸音), 종교 문헌, 의학서·병학서, 조리서, 고소설과 일기류 등 다양한 장르의 자료는 특정 시기의 언어 사용 양상과 사회문화적 맥락을 압축적으로 보여준다. 그러나 이러한 자료들은 비표준적 표기, 필사 과정의 오류, 복잡한 형태 변화로 인해 자동처리와 체계적 디지털 분석이 쉽지 않다. 현대 한국어를 대상으로 한 형태소 분석기는 이미 상당한 성과를 거두었지만, 이를 역사 한국어 자료에 그대로 적용할 경우 분석 오류율이 지나치게 높아 실질적인 연구 활용이 불가능하다. 이는 곧 역사 자료 전용 분석 체계의 필요성을 드러내며, 국어학 및 디지털 인문학 연구가 직면한 새로운 도전 과제라 할 수 있다.

2) 한편, 역사 자료 형태소분석말뭉치 및 형태소 분석 프로그램 개발 연구가 김진해 외(2009)에서 시도된 바 있다. 그러나 고소설에 한정하여 형태소 분석말뭉치가 구축되었고 형태소 분석 프로그램도 그에 맞춰 설계되었다. 이 형태소 분석 프로그램은 기계적으로 고소설의 고빈도 어절 중심으로 구축되어 있어 다른 역사 자료에 적용하기는 한계가 있다. 또한 김미경 외(2016)에서 '형태소 깎는 노인'이라는 형태분석 보조기를 개발한 바 있고, 조은경·한영균(2016)에서는 어휘 분석기 알고리즘 개발을 논의한 바 있으나 그 성과는 전면적인 국어사 자료에 활용하기 어려운 실정이다. 한편, 『17세기 국어사전』, 『표준국어대사전』, 『우리말샘』 등도 말뭉치를 활용한 용례 추출이 활용되기는 하였으나 역사 원시말뭉치를 단순히 검색하여 추출한 자료로 구축된 것이어서 그 활용 면에서는 초기 단계라 할 수 있다.

필요하다.

본 연구팀은 울산대학교 옥철영 교수가 개발한 'UTagger-훈민정음(ver. 0.9)'(데모 프로그램)을 분석·보완하고, 그 성능을 고도화하는 과정을 진행하였다. 'UTagger-훈민정음(ver. 0.9)'은 울산대학교 한국어처리연구실에서 개발한 UTagger3.0을 기반으로 하며, 이 시스템은 수정된 세종 형태 의미 주석 말뭉치를 토대로 구축된 형태소 분석 및 동형이의어 동시 태깅 시스템이다. 특히 기분석 어절 내에서 부분 어절 단위의 결합과 실질형태소·형식형태소의 다양한 조합을 처리할 수 있도록 설계되었다. 테스트용 말뭉치를 대상으로 한 실험에서 학습 말뭉치에 출현하지 않은 어절에 대해 99.06%의 재현율로 형태소를 분석하였고, 기존의 규칙 기반의 형태소 분석기에 비해 빠르면서도(초당 4만 8천 어절) 더 정확한 결과(96.76%)를 보였다.[3] 또한 UTagger3.0은 형태 의미 주석 말뭉치에서 인접 어절 간의 다양한 전이관계 통계를 추출하고, 이를 기반으로 단계별로 전이모델이 적용되는 "단계별 전이모델"을 정의하였다. 이때 각 전이모델에 적용되는 최적의 가중치를 실험을 통해 설정하였다.[4]

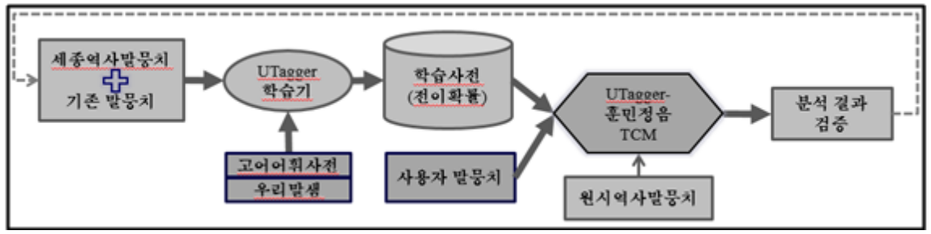
Unicode의 한국어를 형태소 분석하기 위하여 후속으로 개발된 UTagger4.0은 한국어 분석 엔진을 모두 교체하였고, UTagger3.0의 어절간 전이확률(AF, EF, FF) 외에도 좌측어절의 마지막 2개의 형태소(a2), 좌측으로 가장 가까운 실질형태소의 마지막 음절(L1), 좌측 음절 2개와 우측 음절 2개(c1) 등의 전이확률을 활용하였다. 각 전이확률 간의 적용 가중치는 기계학습의 경사하강법을 통해 최적화되었으며 실험 결과 UTagger4.0은 약간 향상된 96.90%의 정확률과 12.0초의 분석 시간을 달성하였다. 세종 형태 의미 말뭉치 중 테스트용의 100만 어절을 대상으로 한 세 종류의 UTagger의 정확률과 분석 시간은 다음 <표 1>과 같다.

모델	정확률	분석시간
UTagger3.0-HMM	96.49%	21.1sec
UTagger3.0-SCP	96.42%	10.0sec
UTAgger4.0-UMA	96.90%	12.0sec

[표 1] UTagger 정확률 및 분석 시간 비교

Unicode 기반의 UTagger4.0은 옛한글(아래아, 합용병서 등) 어절의 형태소 분석 가능성을 열었으며, 이를 바탕으로 역사 자료에 특화된 'UTagger-훈민정음'이 구축되었다. 이 과정에서 기존 세종 역사 말뭉치(2011년 배포, 65개 파일, 약 63만 어절)를 정비하여 학습에 활용하였으며, 한양 PUA 코드를 UTF-8로 변환하고, 방점 및 한자를 제거하는 등 데이터 정규화를 실시하였다. 또한 주석 오류를 수정하고, 학습 과정에서 역사 말뭉치에 10배 가중치를 부여하여 동일 어절 분석 시 역사적 변이가 우선 적용되도록 하였다.

UTagger는 형태 의미 주석 말뭉치로부터 분석에 필요한 통계치를 자동 추출하는 데이터 구동형 구조를 지니므로, 학습용 말뭉치가 확대될수록 기분석 어절이 증가하고, 결과적으로 분석 정확도가 향상되는 선순환 구조를 형성한다(그림 1). 따라서 특정 시대나 장르에 국한되지 않고, 일정 규모의 주석 말뭉치만 확보되면 새로운 분석기가 자동으로 구축될 수 있다. 또한 이를 기반으로 제작된 '사용자 대화형 태깅 도구(TCM)'는 원시 말뭉치 수정, 미학습 어절 판별, 후보 분석 선택, 새로운 분석 결과 추가, 통계 추출, 표준국어대사전 연계 등 다양한 기능을 제공하여, 기존의 전면 수작업 방식에 비해 주석 말뭉치 구축 속도를 현저히 향상시켰다. 특히 사용자 말뭉치 기능을 활용하면, 기존 학습 말뭉치에 없는 어절의 분석 패턴을 즉시 등록할 수 있어 학습 과정 없이도 임시 적용이 가능하다. 이는 반복적으로 출현하는 미학습 어절의 수정 작업을 대폭 줄여, 국어사 형태소 분석 말뭉치 구축의 효율성을 높이는 핵심 요소라 할 수 있다.



[그림] UTagger의 선순환 구조

위 [그림 1]과 같은 구조로 반복 학습된 'UTagger-훈민정음'을 통해 현재 약 60만 어절의 형태소 분석 말뭉치를 구축한 상태이다. 장요한 외(2005)에서 학습된 자료와 미학습된 자료를 통해 어절 및 형태소 경계 인식률과 품사 태깅의 정확도를 살펴본 바 있다. 학습된 『월인석보』 권 7과 미학습된 『월인석보』 권 20의 재현율(recall ratio)의 경우 100%로 나왔고 정확률(precision ratio)은 약간 차이를 보였다. 두 자료가 모두 『월인석보』라는 점에서 『월인석보』 권 20가 완전히 미학습된 자료로 보기는 어렵기는 하고 원자료가 띄어쓰기가 되어 있어서 재현율은 아주 높은 편이다. 정확률은 이 문헌만 보고 평가하기는 어려우나 두 자료 모두 90% 이상으로 나타났다.[5] 문헌마다 차이는 있으나 [그림 1]과 같은 순환 구조로 반복 학습을 지속한다면 상용 가능한 단계에 이를 것으로 보인다.

3) 신준철·옥철영(2012), "기분석 부분 어절 사전을 활용한 한국어 형태소 분석기", 정보과학회논문지:소프트웨어 및 응용 제39권 제5호, 한국정보과학회, 415-424.

4) 신준철·옥철영(2012), "한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델", 정보과학회논문지: 소프트웨어 및 응용 제39권 제11호, 한국정보과학회, 889-901.

5) 장요한외 (2005)에서 평가한 『월인석보』 권 7과 『월인석보』 권 20의 재현율과 정확률은 아래와 같다.

분석 항목	분석 항목 수
재현율	$\frac{5048(B)}{5048(A)} \times 100 = 100\%$
정확률	$\frac{4794(D)}{5048(B)} \times 100 = 94\%$

[표 1] 『월인석보』 권7의 형태 분석 태깅 결과값

분석 항목	분석 항목 수
재현율	$\frac{1760(B)}{1760(A)} \times 100 = 100\%$
정확률	$\frac{1633(D)}{1760(B)} \times 100 = 92\%$

[표 2] 『월인석보』권 20의 형태 분석 태깅 결과값

2.2. 역사 자료 형태소 말뭉치 구축 체제

본 연구팀에서 구축하고 있는 역사 자료 형태소 말뭉치는 다음과 같은 체제를 갖추고 있다.

구분	내용
데이터 수집 및 디지털화	• ‘21세기 세종계획’에서 구축한 원시 말뭉치(이하 ‘세종 말뭉치’) • 형태 분석 말뭉치(이하 ‘세종 형태 말뭉치’) • 날개셋 코드 전환(UTF-8)
형태소 분석 도구	• 역사 한국어 특화 분석기 “UTagger-훈민정음TCM” 활용
형태소 주석	• 어절 단위 분석, 품사 태깅, 원문–분석문 매핑
검증 및 교정	• 전문가 검수와 반자동화 보정 시스템 적용

[표 2] 역사 자료 형태소 말뭉치 구축 체제

■ 데이터 수집 및 디지털화

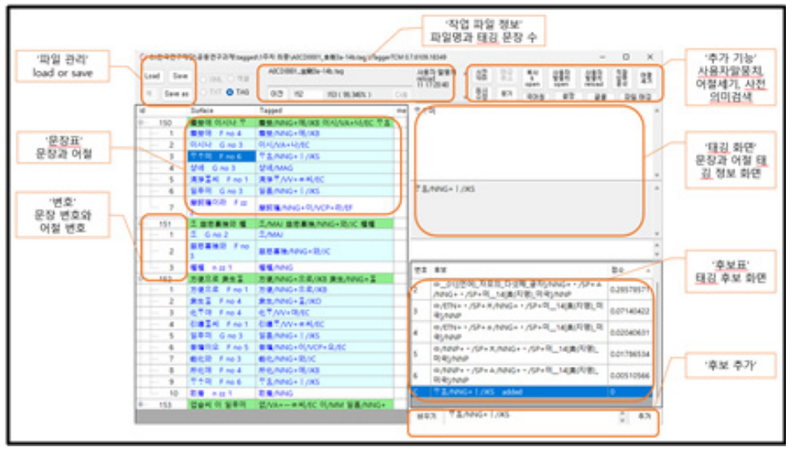
본 연구팀에서는 국립국어원의 ‘모두의 말뭉치’에 공개되어 있는 ‘세종 말뭉치’와 ‘세종 형태 말뭉치’를 신청 후 승인받아 활용하고 있다.[6] ‘세종 말뭉치’는 XML(eXtensibe Markup Language) 파일 1,153 종을 UTF-8로 인코딩한 XML 파일로 제공하고 있는데 본 연구 사업을 시행할 때 받은 파일은 한양 PUA 코드 혹은 UTF-16으로 인코딩된 파일이 있어서 이를 날개셋 전환기를 통해 UTF-8로 전환 처리하였다.[7] ‘UTagger-훈민정음’이 UTF-8로 구조화되어 있어 이를 고려하여 전처리한 것이다.

■ 형태소 분석 도구

본 연구팀에서 자체적으로 개발한 태깅 도구인 ‘UTagger-훈민정음 TCM’을 활용하여 역사 자료 형태소 분석 말뭉치를 구축한다. 이 프로그램은 소규모의 형태소 분석 태깅 작업에 매우 효과적으로 설계된 사용자 대화형 태깅 도구로 학습용 태깅 작업의 효율성과 정확성을 높여 주므로 수작업의 오류를 줄일 뿐 아니라 일관된 형태소 구축이 가능하고 태깅의 시간을 단축할 수 있다. ‘UTagger-훈민정음’ 태깅 내용의 오류를 수정하고 정답 후보의 점수를 조정할 수 있는 태깅 도구이다.
화면 구성은 아래와 같다.

6) 국립국어원의 ‘모두의 말뭉치’에 공개되어 있는 세종 말뭉치는 15세기 한글 창제 이후부터 20세기 초까지 한글로 기록된 문헌자료 1,153종을 UTF-8로 인코딩한 XML 파일로 제공하고 있다. 세종 말뭉치 XML 파일은 원시 말뭉치로 정규식 등으로 검색하고자 하는 형태를 검색하는 데에는 활용할 수 있다. 한편, ‘어듸메’(https://akorn.bab2min.pe.kr)와 ‘kohico’(https://kohico.kr)에서도 내려받을 수 있다.

7) 날개셋(http://moogi.new21.org/ngs_download.htm) 한글 입력기는 Windows용 다용도 한글 입력 프로그램이다.



<그림 > UTagger훈민정음-TCM’의 화면 구성

위 ‘UTagger-훈민정음 TCM’은 ‘파일 관리’(load or save), ‘분석 태깅 정보 화면’과 ‘확인 및 수정 화면’, ‘작업 파일 정보’, ‘추가 기능’으로 구성되어 있다. ‘파일 관리’는 파일을 열거나 저장할 때 사용하는 부분이고, ‘분석 태깅 정보 화면’은 파일을 업로드하면 자동 태깅되어 나온 정보를 확인할 수 있는 화면이다. 이 구성에는 ‘문장표’(문장과 어절 태깅 정보 확인), ‘번호’(태깅된 문장과 어절 번호)로 구성되어 있어서 1차 태깅된 정보 내용이 확인되고 해당 문헌의 문장 및 어절 수도 확인된다. ‘UTF-8’ 형식의 txt 파일을 입력받아 어절 단위 태깅 정보를 확보할 수 있도록 구성되었다.

■ 형태소 주석

텍스트는 우선 문장 단위로 분할하고 문장은 다시 어절 단위로 분할한 뒤 각 어절에 대해 형태소 단위 분석과 품사 태깅을 부여하도록 하였다. 이 과정에서 원문 표기와 분석문이 병렬적으로 제시될 수 있도록 매핑 구조를 설계하였다. 이 형태소 주석은 앞서 소개한 ‘UTagger-훈민정음 TCM’을 통해 1차 수행된다. 이 프로그램에서 태깅 분석 완료된 파일은 ‘.tag’로 자동 저장되는데 그 매핑 구조는 아래와 같다.

(1) 가. 나죄 가 필죵이ᄃᆞ려 모ᄅᆡ 갈 양으로 일오라 ᄒᆞ소.
나. 나죄/NNG 가/VV+ㅏ/EC 필죵/NNP+이/XP+ᄃᆞ려/JKB 모ᄅᆡ/NNG 가/VV+ㄹ/ETM 양/NNB+으로/JKB 일오/VV+라/EC ᄒᆞ/VV+소/EF+./SF

‘.tag’로 저장된 파일은 위 (1)처럼 원문과 분석문의 매핑이 자동으로 실행된다. 텍스트 파일을 문장 단위 분석하여 줄이 나뉘고 문장은 다시 어절 단위로 분석한 뒤 각 어절은 형태에 따라 품사 태깅이 제시된다. 어절은 한 칸 띄어 실현되고 어절 내의 형태들은 ‘+’로 분석되어 구분된다.

■ 검증 및 교정

태깅 도구인 ‘UTagger-훈민정음 TCM’은 사용자 대화형 태깅 도구로 태깅 결과물을 교정 및 검수가 가능한 프로그램이다. ‘UTagger-훈민정음 TCM’에서 자동 분석된 결과물은 아직은 오류율이 높기 때문에 국어학 전공자가 1차 검수를 진행한 뒤에 2차 작업자가 1차 검수 파일을 ‘UTagger-훈민정음 TCM’에 재실행하

여 2차 검증과 교정을 수행한다. 작업 중에 자주 나타난 오분석 내용은 '사용자말뭉치open/reload'에 저장하여 작업자 간에 공유할 수 있도록 한다. 또한 후보표의 점수를 조정하여 정답 후보가 선택될 수 있도록 보정할 수 있다. 최종 3차 검수가 완료된 파일은 'UTagger-훈민정음'에 반복적으로 학습을 진행하여 분석기의 정확도를 점진적으로 향상시켰다. 현재 60만 어절을 통해 6차 학습이 진행된 상태이다.

3. 역사 자료 형태소 말뭉치와 AI 활용

3.1. 역사 자료의 형태 분석 말뭉치의 AI 활용 효과

AI가 인간의 언어로 기록된 방대한 지식을 활용하기 위해서는 우선 자연어 처리(NLP, Natural Language Processing)가 필요하다. 즉 AI가 인간의 언어를 이해할 수 있는 데이터로의 변환과 문법적 구조를 분석하고 의미를 파악할 수 있는 NLP 기술이 필수적이다. 본 연구팀의 'UTagger-훈민정음'은 역사 자료의 원시 말뭉치를 어절 단위로 분할하고 분석된 형태의 품사 정보가 태깅된 파일을 생성하므로 AI에 직접 활용할 수 있는데 앞으로 역사 자료 형태소 말뭉치가 구축하여 이를 AI에 적용한다면 다음과 같은 효과를 기대할 수 있다.

(2) 가. 학문적 가치(데이터 기반 연구)
. 역사 자료의 다양한 표기와 음운 변이, 어휘 및 문법을 학습하게 되면 언어의 변화의 과정을 데이터 기반으로 추적할 수 있다.
나. 디지털 인문학 연구 방법 혁신
. 대규모의 자료의 통계 패턴(빈도 및 공기 관계, 텍스트마이닝 등)을 빠르게 추출할 수 있다.
. 연구자의 직관으로 탐색하기 어려운 대규모 자료에 대한 언어 사용 규칙이나 의미 네트워크를 드러낼 수 있다.
. 자동화된 도구 개발: 역사 자료 한글의 광학문자인식(OCR, Optical Character Recognition) 및 자동 번역기를 개발할 수 있다.
다. 교육 및 문화 활용
. 중세 및 근대 국어 자료의 주석과 현대어 자동 변환 등을 통해 교육용 자료 활용 가능하다.
. 고문헌, 고소설, 비문과 한글 편지 등을 현대어 변환을 자동화하면 문화 콘텐츠 활용 가능성을 높일 수 있으며 영화나 드라마, 게임, 출판 등의 문화 산업에도 활용될 수 있다.
라. AI의 언어 능력 향상
. 역사 자료를 통해 언어의 역사성과 다양성이 반영되어 모델의 언어 능력을 확장할 수 있다.
. 옛한글을 바탕으로 한 문학적 문체 생성 및 역사적 맥락 재현 같은 새로운 응용도 가능할 수 있다.

3.2. OpenAI 활용

이 절에서는 'UTagger-훈민정음'을 통해 구축한 『순천김씨 묘 출토 간찰』의 형태소 분석 말뭉치를 OpenAI ChatGPT(GPT-5)를 활용하여 공기어 분석과 네크워크 시각화 및 중심성 분석, 주제별 군집 분류 및 히트맵 시각화를 시도하여 텍스트의 특성을 밝히고자 한다.[8] 인문 디지털 연구 방법에서 공기어 분석은 단어 간 연관성과 주제적 유사성을 파악할 수 있는 유용한 도구로 텍스트 전체의 특성을 규명하는 데 활용된다. 더

8) 『순천김씨 묘 출토 간찰』의 한글 편지는 발신자와 그 성별이 분명하기 때문에 발신자의 성별에 따라 편지를 구분하기에 매우 유용하고 언간의 특성상 동일한 발신자라 하더라도 수신자에 따라 언어 표현이 달라지기 때문에 발신자와 수신자의 관계에 따라 편지를 비교 관찰하는 것이 흥미로운 연구 주제이다. 이와 관련하여 박주자(2018), 이해로(2019), 장요한(2005) 등에서 이 발신자와 수신자의 성별에 따라 언어적 양상을 검토한 바 있다. 여기에서는 이 편지들을 OpenAI를 통해 인문 디지털 연구 방법을 시도해 보고자 한다.

나아가 시대별 말뭉치를 비교하면 공기어 네트워크의 변화를 통해 단어의 의미 변화와 사용 양상의 변천을 추적할 수 있다.

중심성 분석은 네트워크 내에서 각 단어가 점유하는 구조적 위치를 파악하여 특정 단어가 텍스트에서 어떤 영향력을 가지는지를 밝히는 방법이다. 이 방법은 특정 시점에서 단어의 중요도를 확인할 뿐 아니라, 시간에 따른 중심성 변화를 통해 단어의 영향력 추이를 분석하는 데에도 활용될 수 있다. 한편, 주제별 군집 분석은 공기어를 기반으로 유사한 의미의 단어를 자동으로 군집화하여 텍스트의 주제를 탐색하고 사회 계층이나 지역 방언에 따른 언어 사용 차이를 연구하는 데 활용될 수 있다.

그간의 국어사 연구에서 역사 자료 말뭉치는 주로 어휘 및 형태 검색, 빈도 조사 등에 머물렀고 디지털 연구에는 활용되지 못하였다. 주시하는 바와 같이 역사 자료를 인문 디지털 연구에 활용하려면 우선적으로 NLP 기반 분석이 필요하나 주석 말뭉치의 성과가 아직은 제한적이어서 연구자가 적극적으로 활용하기가 쉽지 않은 상황이다. 이에 본 절에서는 'UTagger-훈민정음'을 통해 구축한 『순천김씨 묘 출토 간찰』의 형태소 태깅 말뭉치를 발신자와 수신자의 관계에 따라 전체 코퍼스를 1그룹(남성>여성), 2그룹(여성>여성), 3그룹(여성>남성)으로 구분하고 인문 디지털 방법을 도입하여 그 결과를 도출하고자 한다.[9][10] 이는 발신자와 수신자의 관계에 따른 언어 사용 패턴과 텍스트 구성의 차이를 살펴보고자 하는 것이다.[11] 이 연구는 언어 연구의 범위를 확장하기 위한 방법론적 시도라 하겠다.

3.2.1. 남성이 여성에게 보낸 편지 분석

1그룹(남성>여성)으로 분류된 형태소 분석 말뭉치는 '남_여.tag' 파일로 저장되어 있고 이 파일을 OpenAI ChatGPT(GPT-5) 업로드하여 명사류(NNG, NNP, NNB, NR)에 대한 공기어 분석과 공기어 네크워크 시각화 및 중심성 분석, 주제별 군집 분류 및 히트맵 시각화를 진행하였다.[12][13]

9) 이 작업에서 구축한 『순천김씨 묘 출토 간찰』의 형태소 말뭉치는 몇 가지 전치리를 수행하였는데 상세한 전처리 과정은 장요한(2005)에서 제시한 것으로 자세한 전처리 과정을 논문을 참고하기 바란다.

10) 남성 편지에 나타난 형태의 총 유형(type)은 1,361개이며 총 빈도(token)는 6,632개이다. 한편 여성 편지에 나타난 형태의 유형은 2,677개이며 총 빈도는 22,956개이다.

11) 주지하는 바와 같이 『순천김씨 묘 출토 간찰』의 한글 편지 189 편은 16세기 자료로 알려진 후기 중세 국어 자료이고 발신자와 수신자가 분명하지 않은 6 편, 한문과 이두가 중심인 2편을 제외한 181 편은 발신자와 수신자가 분명하다(조항범, 1998a:8-9). 발신자는 순천(順天) 김씨(金氏) 친정 어머니인 신천(信川) 강씨(康氏), 친정 아버지 김훈(金訓), 순천 김씨의 남편 채무이(蔡無易), 남동생 김여흘(金汝屹)과 김여율(金汝律)이고 수신자는 순천 김씨 외에도 그의 남동생과 여동생, 남편, 올케 등이다. 이 편지들은 발신자와 그 성별이 분명하기 때문에 발신자의 성별에 따라 편지를 구분하기에 매우 유용하다. 물론 언간의 특성상 동일한 발신자라 하더라도 수신자에 따라 언어 표현이 달라지겠지만 계량적 지표에서 이러한 사항은 확인될 수 있을 것이다.

12) 다양한 조건의 공기어 분석을 시도하면 풍부한 분석 자료를 확인할 수 있겠지만 여기에서는 제한된 지면의 문제가 있어 명사류를 중심으로 다루고자 한다. 장요한(2005)에서는 빈도 조사 내용이기는 하지만 품사와 어미에 따라서 그 빈도의 차이를 검토한 바 있다.

13) OpenAI, ChatGPT (GPT-5), https://chat.openai.com/ (접속일: 2025.9.27).

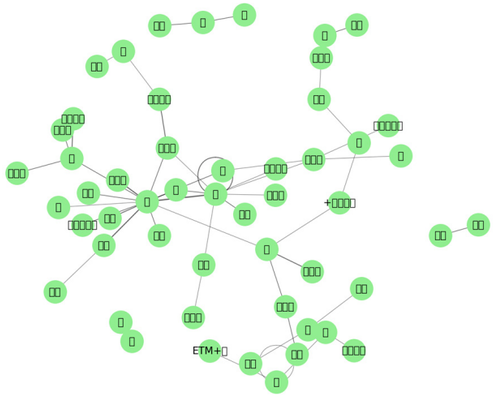
■ 공기어 분석

상위	1그룹(남>여)	
	공기어 쌍	빈도
1	것, 나	6
2	나, 자내	5
3	쁟, 나	5
4	것, 것	4
5	나, 말	4
6	늬일, 오늘	4
7	나, 유무	4
8	민, 채	4
9	늠, 집	4
10	녀름, 일	4
11	것, 나죄	3
12	늬일, 나	3
13	나, 보기	3
14	날, 편지	3
15	나, 일	3
16	민셔방, 안해	3
17	믁숨, 나	3
18	민셔방, 집	3
19	무명, 필	3
20	즉식, 일	3

[표 3] 1그룹(남>여) 공기어 순위

제1 그룹에서는 위 [표 3]에서와 같이 '나-것'(6회), '나-자내'(5회) 등 화자와 청자의 관계를 드러내는 쌍이 반복적으로 나타났다. 또한 '쁟-나'(5회), '나-말'(4회)와 같은 경우는 화자가 자신을 중심으로 정보를 전달하거나 인지하는 행위를 강조하는 표현으로 볼 수 있다. '것-것'(4회)의 반복은 발화의 내용을 전달하는 특징을 보여준다.

■ 네크워크 시각화 및 중심성

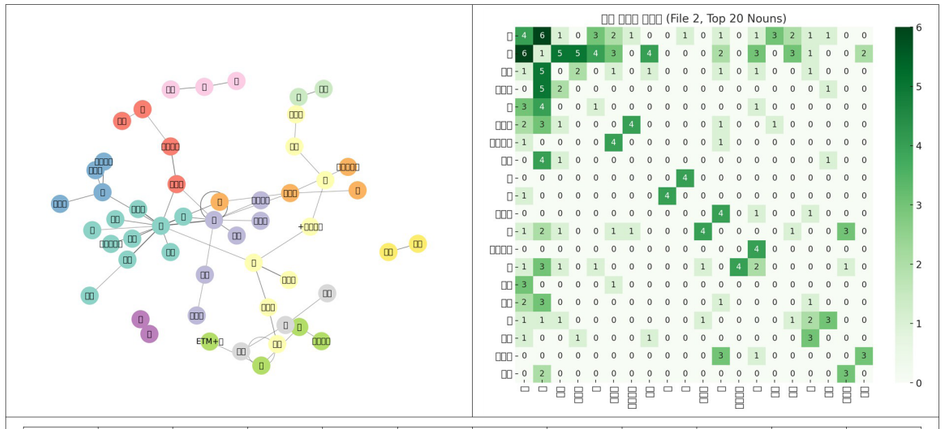


<그림 > 공기어 네트워크 시각화와 중심성

네트워크 시각화하는 과정에서 옛한글이 깨지는 현상이 있으나 분석 결과 '나', '것', '집', '죵', '일'이 중심성을

보였다. 특히 '나'가 가장 높은 연결 중심성(Degree)과 매개 중심성(Betweenness)을 보인 것은 언간 텍스트의 성격상 화자인 '나'가 핵심 주제어이면서 의미 연결 고리 역할을 수행하고 있기 때문인 것으로 해석된다. '집'과 '죵', '일'은 생활 공간과 사회적 관계, 구체적 사건을 반영하며 개인과 삶의 관계망이 텍스트에 드러난 것으로 이해할 수 있다. 즉, 이 텍스트에서는 화자(나)를 중심으로 대상이나 내용 지시(것)와 생활 공간(집), 사회적 관계(죵), 구체적 사건(일)이 서로 연결된 구조가 확인된다.

■ 주제별 군집 분류 및 히트맵 시각화



community1	community2	community3	community4	community5	community6	community7	community8	community9	community10	community11	community12
말	민셔방	복거리	밤	즉석	제	ㅣ	필	아모	채	편지	만장
쁟	홍	글사니	오늘	녀름	물	거줄	반	빼	민	날	이개
자내	싱원	것	댱명	즉식	예	병	무명	듸			
보기	이제	사롬	늬일	일	안싱원	죵					
믁숨	늠	필죵									
부모	손쳠디	나죄									
글	안해										
유무	집										
나											
어제											

[표 4] 주제별 군집 분류 및 히트맵 시각화

단어들 간의 네트워크를 모듈성 기반으로 분류한 결과, 12개의 작은 그룹(Community)이 도출되었다. 그 중에서 'Community 1'은 '말, 쁟, 자내, 믁숨 …'과 같이 화자와 청자, 발화 내용으로 구성되어 있고 'Community 2'는 '민셔방, 싱원, 늠, 손쳠디'과 같이 가족의 관계를 중심으로 이루어져 있고 'Community 4'는 '밥, 오늘, 늬일' 등과 같이 일상의 식사와 시간으로 구성되어 있다. 시각화된 동시출현 빈도는 '나-것', '나-자내', '나-말'에서 높게 나타났다. 이를 통해 이 텍스트가 화자(나) 와 청자(자내) 관계를 축으로, 발화내용(말, 쁟)과 생활 단위(밥, 오늘, 집)까지 긴밀히 얽혀 있음을 짐작할 수 있다.

3.2.2. 여성이 여성에게 보낸 편지

2그룹(여성>여성)으로 분류된 형태소 분석 말뭉치는 '여_여.tag' 파일로 저장되어 있고 이 파일을 OpenAI ChatGPT(GPT-5) 업로드하여 명사류(NNG, NNP, NNB, NR)에 대한 공기어 분석과 공기어 네크워크 시각화 및 중심성 분석, 주제별 군집 분류 및 히트맵 시각화를 진행하였다.

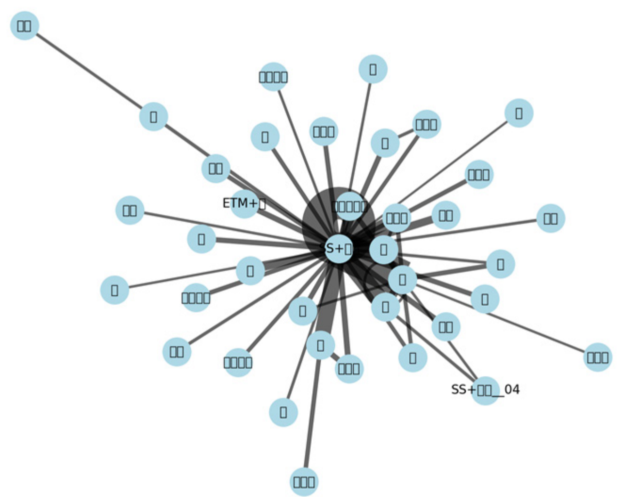
■ 공기어 분석

상위	2그룹(여>여)	
	공기어 쌍	빈도
1	순천, 나	151
2	순천, 순천	129
3	순천, 것	73
4	순천, 일	72
5	순천, 너	64
6	순천, ᄌᆞ식	47
7	것, 나	34
8	순천, 우리	31
9	순천, 말	29
10	순천, ᄆᆞᅀᆞᆷ	27
11	순천, 아바님	23
12	나, 일	22
13	너, 아바님	22
14	순천, 유무	22
15	ㅣ, 순천	22
16	순천, 집	21
17	순천, 제	21
18	ᄆᆞᅀᆞᆷ, 나	20
19	순천, 적	20
20	순천, 종	20
21	순천, 너희	19
22	나, 몸	19
23	너, 오라비	18
24	순천, 아히	18
25	순천, 요ᄉᆞ이	18
26	순천, 긔별	18
27	ᄌᆞ식, 나	17
28	순천, 내	17
29	순천, 싱원	16
30	순천, 병	16

[표 5] 그룹 2(여>여) 공기어 순위

여성 간 편지의 경우는 ‘순천’이 압도적으로 중심이 되는 공기어로 나타났다. 상위 공기어 관계를 보면 ‘순천-나’, ‘순천-것’, ‘순천-일’, ‘순천-ᄌᆞ식’, ‘나-일’은 특정 지역과 화자(나), 업무(일)을 반영하고 있으며 ‘순천-ᄆᆞᅀᆞᆷ’, ‘순천-우리’, ‘ᄆᆞᅀᆞᆷ-나’, ‘순천-말’, ‘순천-아바님’, ‘너-아바님’, ‘너-오라비’ 등은 감정 표현과 가족 관계를 나타내고 있다. 한편, ‘유무, 병, 종, 긔별, 몸’ 등은 안부와 건강을 나타내는 주요 주제어로 이 텍스트가 순천이라는 공간적 배경을 중심으로 화자와 업무(일), 가족 관계와 건강 상태를 주된 관심사로 다루고 있음을 알 수 있다.

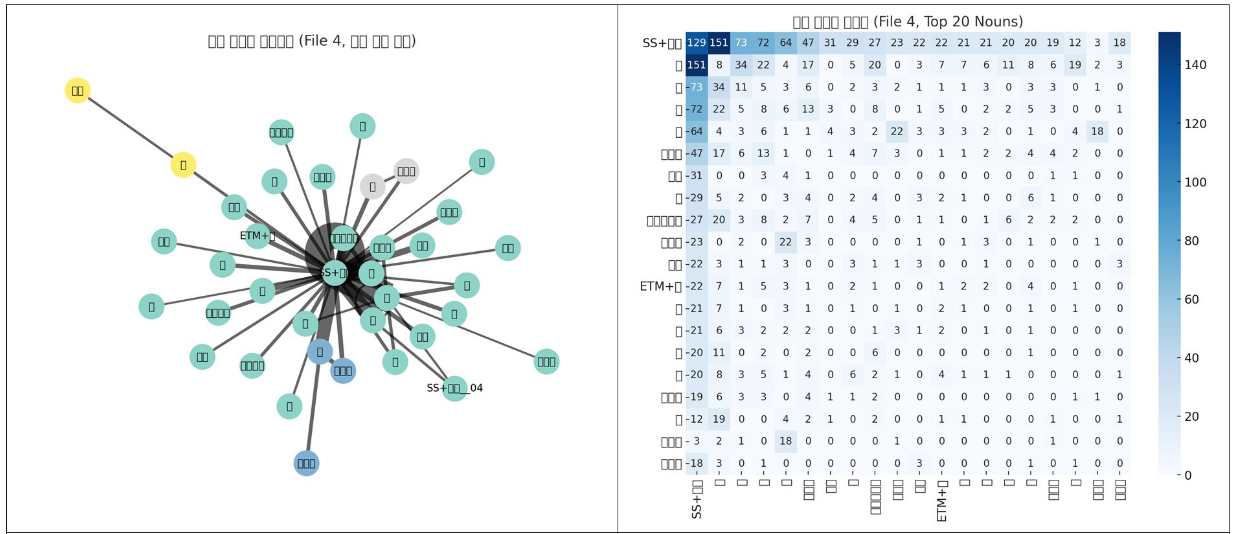
■ 네크워크 시각화 및 중심성



<그림 > 네트워크 시각화 및 중심성

네크워크 시각화 결과 가장 중심적인 단어는 “순천”이며 그 다음으로 “나, 너, 것, 일, 집” 등이 높은 연결성을 보였다. 즉, 지역(순천)과 인물 관계(나–너), 사건/사물(일, 것, 집)이 긴밀하게 얽혀 텍스크가 전개되어 있음을 짐작할 수 있다.

■ 주제별 군집 분류 및 히트맵 시각화



community1	community2	community3	community4
이제	아바님	채셔방	필
말	너	집	무명
아기	오라비		
것			
나			
적			
넌			
너희			
긔온			
날			
이			
제			
일			
ᄆᆞᅀᆞᆷ			
믈			
아히			
긔별			
순천			

[표 6] 주제별 군집 분류 및 히트맵 시각화

단어들 간의 네트워크를 모듈성 기반으로 분류한 결과, 4개의 작은 그룹(Community)이 도출되었다. ‘Community 1’은 ‘이제, 말, 아기, 것, 나’ 등과 같이 화자와 발화, 그 가족 관계로 구성되어 있고 ‘Community 2’는 ‘아바님, 너, 오라비’ 등과 같이 가족의 인물 관계를 중심으로 묶였다. ‘Community 3’은 ‘집’을 중심으로 한 생활 공간 관련 군집이며 ‘Community 4’는 ‘무명’과 같이 사물 중심의 군집이다. 이를 통해서 이 텍스트가 순천이라는 공간을 중심으로, 화자와 청자(나–너) 그리고 가족 관계(아바님, 오라비, 아기), 생활 공간(집)과 일(무명)이 서로 긴밀히 연결되어 있음을 알 수 있다. 즉 지역과 가족, 생활 공간, 생활 공간에서의 일이 핵심 주제로 파악된다.

3.2.3. 여성이 남성에게 보낸 편지

3그룹(여성>남성)으로 분류된 형태소 분석 말뭉치는 ‘여_남.tag’ 파일로 저장되어 있고 이 파일을 OpenAI

ChatGPT(GPT-5) 업로드하여 명사류(NNG, NNP, NNB, NR)에 대한 공기어 분석과 공기어 네크워크 시각화 및 중심성 분석, 주제별 군집 분류 및 히트맵 시각화를 진행하였다.

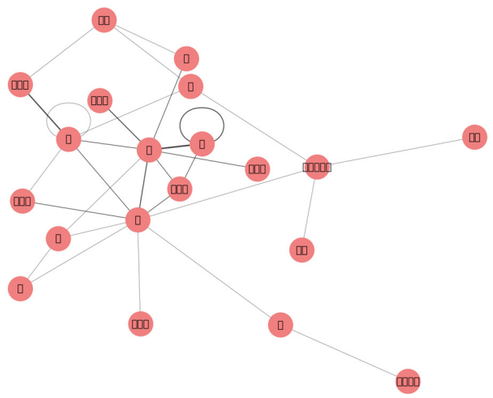
■ 공기어 분석

상위	3그룹(여>남)	
	공기어 쌍	빈도
1	나, 일	5
2	너, 아바님	4
3	것, 나	3
4	뜯, 나	3
5	일, 일	3
6	것, 너	2
7	나, 너	2
8	나, 말	2
9	그듸, 나	2
10	즈식, 것	2
11	그듸, 일	2
12	것, 그듸	2
13	스실, 나	2
14	무슴, 것	1
15	무슴, 이번	1
16	무슴, 우리	1
17	것, 죵	1
18	죵, 하	1
19	것, 하	1
20	나, 하	1

[표 7] 그룹 3(여>남)의 공기어 순위

그룹 3(여>남)의 경우는 '나-일'(5회), '너-아바님'(4회), '것-나'(3회), '뜯-나'(3회), '일-일'(3회) 등의 공기어 쌍이 상위 순위를 차지하였다. 화자 개인 경험과 사건이 연결되어 있고 청자와 부친 간의 관계도 중요한 의미망을 형성하고 있음을 보여준다.

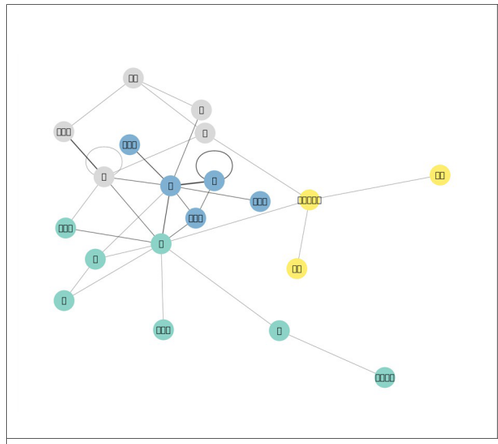
■ 네크워크 시각화 및 중심성

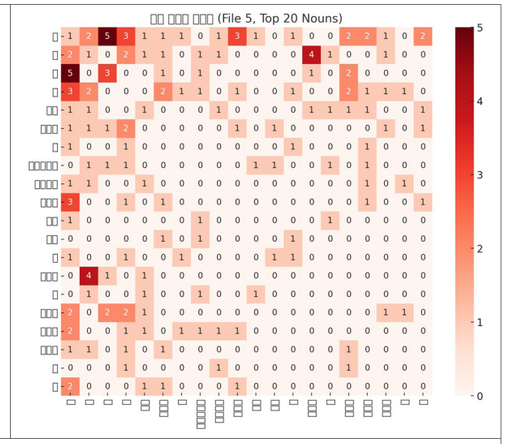


<그림 > 네트워크 시각화 및 중심성

네트워크 시각화 결과, 시각화에서 '나', '너', '것', '일', '무슴' 등이 주요 노드로 나타났다. 즉 화자와 청자, 지시 내용 및 사건, 내적 상태가 서로 연결된 구조를 이루고 있음을 의미한다. 따라서 이 텍스트는 개인 경험과 관계, 감정이 얽힌 내용으로 구성되었다 하겠다.

■ 주제별 군집 분류 및 히트맵 시각화





community1	community2	community3	community4
죵	그듸	녀	우리
사롬	뜯	말	이번
그이	일	제	무슴
것	스실	유무	
즈식	나	아바님	
년			
하			

[표 8] 주제별 군집 분류 및 히트맵 시각화

단어들 간의 네트워크를 모듈성 기반으로 분류한 결과, 4개의 작은 그룹(Community)이 도출되었다. 'Community 1'은 '죵, 사롬, 그이, 것, 즈식' 등과 같이 사회 관계를 중심으로 군집되었고 'Community 2'는 '그대, 뜯, 일, 스실, 나' 등과 같이 청자와 발화 내용, 화자로 구성되어 있다. 'Community 3'은 '너, 말, 유무, 아바님' 등과 같이 대화·호칭·가족 관계로, 'Community 4'는 '우리, 이번, 무슴' 등과 같이 가족 공동체와 시점, 정서 표현으로 구성되어 있다. 동시출현 빈도 역시 '나-것', '너-아바님', '나-무슴'이 높은 빈도로 나타났다. 화자 경험·가족 관계·정서 표현이 핵심 주제임을 보여준다.

3.2.4. 정리

세 그룹은 화자 중심 구조로 전개되는 양상을 보인다. 세 그룹 모두 '나'(화자)가 핵심 노드로 나타났으며 텍스트가 화자 시점에서 전개되는 특징을 보인다. '나-것', '나-말'과 같은 화자 지향적 표현이 반복적으로 나타나는 것에서도 알 수 있다. 또한 세 그룹의 편지가 주제별 군집 구조에서도 유사성을 보인다. 화자와 청자 관계, 가족 관계, 생활 공간 및 일상 사건, 정서 및 감정 등의 범주로 묶이는 경향을 보인다. 이 구조는 세 그룹 모두 유사하다.

그런데 핵심어와 텍스트 측면에서 볼 때 다소 차이를 보인다. 제1 그룹(남성>여성)은 '나', '것', '집', '죵', '일'과 같이 화자와 생활, 사회적 관계가 중심어를 이루며, 화자를 중심으로 한 일상과 업무, 지시적 내용이 두드러진다. 제2 그룹(여성>여성)은 '순천'이 압도적으로 중심어로 작용한다. 지역과 공간성을 바탕으로 가족 관계와 감정 표현이 결합되어 있다. 이 경우 텍스트는 공간과 가족을 중심으로 조직되며, 일상 기록과 안부 및 건강 표현이 특히 강조된다. 제3 그룹(여성>남성)은 '나', 'ᄲᅳᆮ', '아바님', 'ᄆᆞᅀᆞᆷ'과 같이 화자 경험과 가족 관계, 정서 표현이 강조되었다는 점에서 앞의 두 그룹과 차이를 보인다.

군집 특성 측면에서 볼 때 제1 그룹(남성>여성)은 군집 수가 많고(12개) 세분화된 주제를 구성하고 있어서 다양한 관계와 사건 서술되어 있다. 제2 그룹(여성>여성)은 군집 수가 적고(4개) 지역과 가족 중심으로 간결하게 조직되었다. 제3 그룹(여성>남성)은 군집 수가 적고(4개) 화자와 청자, 가족 관계, 정서의 축으로 구성되었다.

종합하면, 세 그룹 모두 화자 중심으로 가족 관계를 강조하고 생활과 정서가 얽힌 구조라는 점에서 공통점을 보이면서도, 제1 그룹(남성>여성)은 생활과 업무 중심의 실질적 내용을, 제2 그룹(여성>여성)은 지역과 가족 중심의 공동체적 내용을, 제3 그룹(여성>남성)은 화자와 청자 관계와 정서적 교류 내용을 중심으로 구성하고 있다는 점에서 차별적 특성을 보인다.

4. 결론

본 연구는 한국어 역사 자료를 대상으로 한 전용 형태소 분석기 'UTagger-훈민정음'의 고도화 과정과 이를 활용한 형태소 분석 말뭉치 구축, 그리고 인공지능 기술과의 융합 가능성을 탐색하였다. 연구 과정에서 기존 현대어 중심 분석기의 한계를 극복하고, 중세 및 근대 한국어 자료의 특수한 음운·형태적 변이를 효과적으로 처리할 수 있는 분석 체제를 마련하였으며, 약 60만 어절 규모의 형태소 분석 말뭉치를 확보하였다.

특히 'UTagger-훈민정음'은 반복 학습 구조와 사용자 대화형 태깅 도구를 통해 분석 정확도를 향상시키고 주석 작업의 효율성을 크게 높였다. 이러한 결과는 국어사 연구에서 비표준 표기와 다양한 언어 변이를 자동 처리할 수 있는 기반을 제공하며, 더 나아가 디지털 인문학적 방법론의 새로운 가능성을 제시한다. 또한 OpenAI 기반 분석 사례를 통해 역사 자료 말뭉치가 공기어 분석, 네트워크 중심성 분석, 주제별 군집화 등 다양한 인문 디지털 연구 방법론과 결합될 수 있음을 확인하였다.

그러나 본 연구는 여전히 자료의 범위와 장르적 한계, 그리고 일부 희귀 어형 처리에 있어 한계를 지닌다. 향후에는 더 방대한 주석 말뭉치를 확보하고, 장르와 시기를 확장하는 동시에 최신 대규모 언어모델과의 연동을 강화함으로써 분석기의 성능을 고도화할 필요가 있다.

정리하면, 본 연구는 과거 한국어 자료를 디지털 언어 자원으로 전환하여 AI 기술과 접목할 수 있는 구체적 가능성을 보여주었다. 이는 국어학과 역사언어학의 연구 방법을 혁신할 뿐 아니라, 교육·문화·산업적 활용으로 이어질 수 있는 잠재력을 지니며, 향후 한국어 역사 자료 연구의 새로운 전기를 마련할 것으로 기대된다.

주제어: 한국어 역사 자료, 형태소 분석기, , UTagger-훈민정음, 형태소 분석 말뭉치 구축, 디지털 인문학, 인공지능, OpenAI ChatGPT(GPT-5), 순천 김씨 묘 출토 간찰

참고문헌

강정희(1987). 「여성어의 한 유형에 관한 연구」, 『국어학신연구』, 탑출판사
김선희·이석규(1992). 「남성어 여성어에 관한 연구」, 『어문학연구』2. 목원대학교 어문학연구, 35-64.
구혜승(2024), 「한국어 에세이 자연어 처리를 위한 한국어 형태소 분석기 개선 방안 연구」, 한국교원대학교 석사학위논문.
국립국어원(2007), 「21세기 세종계획 백서」, 국립국어원.
국립국어원(2021), 「'국립국어원 신문 말뭉치(버전 2.0)' 설명 자료」, 국립국어원.
김동혁 외(2003), 「대용량 문서 데이터베이스를 위한 효율적인 점진적 문서 클러스터링 기법」, 『정보처리학회논문지D』28, 한국정보처리학회, 57-66.
김미경 외(2016), 「형태소 깎는 노인: 국어사 자료를 위한 형태분석 보조기」, 『제28회 한글 및 한국어 정보처리 학술대회 논문집』, 한국어정보학회, 39-43.
김수연, 안석호, 김동현, 이의종, 서영덕(2022), 「형태소 분석기의 품사별 정확성 분석」, 『한국정보기술학회 하계 종합학술대회 논문집』, 한국정보기술학회, 378–381.
김아연(2020), 「형태주석 코퍼스 구축을 활용한 『대한매일신보』 시평가사의 키워드 추출과 분석: 1907년 국문판을 중심으로」. 『역사학연구』 78, 호남사학회, 77-111.
김재한·안미정·옥철영(1993), 「활용 형태소에 기반한 한국어 형태소 분석기」, 『한국정보과학회 학술발표논문집 20(2)』, 한국정보과학회, 1139-1142.
김재한·옥철영(1994), 「통합형태소를 이용한 한국어 형태소 분석기」, 『한국정보과학회 학술발표 논문집 21.2A』, 한국정보과학회, 653-656.
김준수·최호섭·이왕우·옥철영(2002), 「한국어 동형이의어 태깅 시스템 구현」, 『한국정보과학회 언어공학연구회 학술발표 논문집』, 한국정보과학회, 24-30.
김진해(2008), 「활자본 고소설 말뭉치 구축의 국어정보학적 의의: 형태 분석 및 통합 검색 시스템 구축을 중심으로」, 『국어국문학』 149, 국어국문학회, 69-107.
김진해(2023), 「개화기 국어 말뭉치 구축 현황과 개선 방안」, 『2022 겨울 국어사학회 전국 학술대회 발표집』, 국어사학회, 141-162.
김진해·차재은·김건희·이의철(2009), 「歷史資料 형태분석 프로그램 개발의 國語學的 意義와 活用 硏究: 活字本 古小說을 중심으로」, 『어문연구』 37-4, 한국어문교육연구회, 137-162.
김한샘·장연지·강예지(2020), 「통시 말뭉치에 기반한 언어 변화 연구: 20세기 신문 말뭉치의 구축과 분석」, 『한글』 330, 한글학회, 909-947.
김흥규 외(2007), 「21세기 세종계획 국어 기초자료 구축(연구보고서)」, 국립국어원.
남윤진·옥철영(1996), 「말뭉치 분석에 기반한 명사파생접미사의 사전정보 구축」, 『정보과학회논문지(B)』 23.4, 한국정보과학회, 389-401.
도재학·송인재(2021), 「빈도와 공기어 분석으로 본 근대 한국과 중국의 '문명' 개념」, 『어문논집』 31, 민족어문학회, 277-304.
방진우(2020), 「역사 자료 형태분석에서 미등록어 추정과 분석 중의성 해소」, 『동양예학』 44, 동양예학회, 169-197.
신준철·옥철영(2012), 「기분석 부분 어절 사전을 활용한 한국어 형태소 분석기」, 『정보과학회논문지 : 소프트웨어 및 응용』 39(5), 한국정보과학회, 415-424.
신준철·옥철영(2012), 「한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델」, 『정보과학회 논문지 : 소프트웨어 및 응용』 39(11), 한국정보과학회, 889-901.
신준철·옥철영(2014), 「부분어절 조건부확률 기반 동형이의어 태깅 모델」, 『정보과학회논문지: 소프트웨어 및 데이터 공학』 제3권 제10호, 한국정보과학회, 407-420.
신준철·옥철영(2015), 「한국어 어휘의미망(UWordMap)을 이용한 동형이의어 의미분별 정확률 개선에 관한 연구」, 『한국정보과학회 학술발표논문집』, 한국정보과학회, 817-819.
신준철·옥철영(2016), 「한국어 어휘의미망(UWordMap)을 이용한 동형이의어 분별 개선」, 『정보과학회논문지』 43(1), 한국정보과학회, 71-79.
신중진(2004), 「개화기 한글자료 말뭉치의 구축 방안」, 『관악어문연구』 29, 서울대학교 국어국문학과, 261-283.
안예리·이주현(2014), 「20세기 문어 말뭉치 구축을 위한 기초 연구」, 『한국어학』 63, 한국어학회, 229-265.
양승현·김영섬(2000), 「부분 어절의 기분석에 기반한 고속 한국어 형태소 분석 방법」, 『정보과학회논문지 : 소프트웨어 및 응용』

27(3), 한국정보과학회, 290-301.
윤영민(2018), 「말뭉치 구축·활용의 흐름과 현재의 동향: 일본의 사례를 중심으로」, 『언어사실과 관점』 45, 연세대학교 언어정보연구원, 35-59.
이래호(2019), 「<송규렴가 언간>에 나타나는 남녀 간 언어 차이」, 『어문논집』, 중앙어문학회, 7-37.
이래호(2023), 「근대 국어 말뭉치 구축 현황과 개선 방안」, 『2022 겨울 국어사학회 전국 학술대회 발표집』, 국어사학회, 109-140.
이민우(2021), 「의미 변화의 양적 추정: 말뭉치를 이용한 의미 변화 연구」, 『한국어의미학』73, 한국어의미학회, 73-59.
이영제·강범모(2014), 「현대국어 역사 코퍼스를 이용한 언어 변화의 계량적 연구」, 『한국어학』 63, 한국어학회, 267-303.
장경준(2013), 「석독구결 자료의 주석말뭉치 구축에 대하여」, 『한국학연구』 46, 고려대학교 한국학연구소, 167-195.
장경준(2019), 「역사 자료 말뭉치의 보완과 활용: 언해 자료를 중심으로」, 『한말연구』 53, 한말연구학회, 183-218.
장경준 외(2019), 「국어 역사 말뭉치 구축 중장기 계획 수립(연구보고서)」, 국립국어원.
장요한 외(2005), 「역사 자료 형태소 분석 말뭉치 프로그램 개발 및 고도화」, 『언어와 정보 사회』 31, 서강대학교 언어정보연구소, 191-219.
장요한(2005b), 「역사 자료 형태소 분석 프로그램 개발 및 활용 - 『순천김씨 묘 출토 간찰』을 중심으로」, 『언어과학』 32-2, 한국언어과학회, 113-136.
조은경·한영균(2016), 「역사 자료 어휘 분석기를 위한 어휘부 개발 방법」, 『한국사전학』 27, 한국사전학회, 49-74.
조항범(1998a), 『순천김씨 묘 출토 간찰』, 태학사.
조항범(1998b), 「<順天金氏 墓 出土 簡札>에 대한 몇 가지 문제」, 『개신어문연구』15, 개신어문학회, 105-131.
하정수(2023), 「중세 국어 말뭉치 구축 현황과 개선 방안」, 『2022 겨울 국어사학회 전국 학술대회 발표집』, 국어사학회, 91-108.
허인영(2023), 「국어사 말뭉치 활용 현황과 방안」, 『2022 겨울 국어사학회 전국 학술대회 발표집』, 국어사학회, 163-186.
홍윤표(1998), 「국어학 자료의 전산화 방법과 그 학문적 의의」, 『국어국문학』 121, 국어국문학회, 307-326.
홍윤표(2001a), 「한국어 전자 자료의 수집과 정리 및 활용 방안」, 『새국어생활』 11-2, 국립국어원, 37-75.
홍윤표(2001b), 「국어사 자료 코퍼스의 구축 현황과 과제」, 『한국어학』 14, 한국어학회, 1-32.
황문환(2010), 「조선시대 언간 자료의 현황과 특성」, 『국어사연구』, 국어사학회, 73-131.
Lakoff, R. 1975. Language and woman's Place. New York Harper and Row Publishers Inc.

한국어 교육에서의 인공지능 기반 의미 분석: 터키 대학 맥락에서의 시사점

AI-Supported Semantic Analysis in Korean Language Education: Insights from the Turkish University Context

무함메트 에므레 코르크마즈
앙카라대학교 교수

Muhammet Emre Korkmaz
Professor, Ankara University

Abstract

Despite growing research on AI in language education, semantic analysis in Korean as a Foreign Language (KFL) remains underexplored. This study examines the integration of AI-supported semantic analysis in an undergraduate Korean translation course at Ankara University (Spring, 2024-2025, n=11 final-year students). The course combined semantic theory with translation practice across diverse text types, using AI prompts for comparative evaluation and semantic exploration. Student reflections revealed improvements in translation skills (8/11 students), AI literacy (6/11), and vocabulary expansion (5/11). Sentiment analysis showed predominantly positive perceptions (10/11 students). The instructor employed an 8-step iterative workflow balancing AI efficiency with pedagogical quality. Findings indicate that students evolved from over-reliance on AI outputs to more critical, reflective use. This exploratory case study represents the first systematic integration of AI-supported semantic analysis in Turkish KFL contexts, offering a replicable model for under-resourced language programs and establishing a reference point for future curricular innovation.

Keywords: artificial intelligence, Korean language education, semantic analysis, translation training, Turkish KFL learners

1. Introduction

The rapid integration of generative artificial intelligence (AI) into everyday life has heightened academic and professional interest in applying these tools across diverse domains, including foreign language education. While AI-supported pedagogical approaches have begun to draw attention in certain linguistic contexts, their role in Korean as a Foreign Language (KFL) education is only just beginning to emerge as a research focus.

In recent years, interest in learning Korean as a foreign language in Türkiye has expanded considerably, shaped by both cultural influences and institutional developments. The three Korean Language and Literature departments at Turkish universities constitute the primary institutional framework for KFL education and Korean Studies research. Despite this expansion, disparities among curricula and the lack of standardized approaches remain evident.

Against this backdrop, the Comparative Semantic Research course (KRD422 Karşılaştırmalı Anlambilim Araştırmaları, 대조의미론 연구) at Ankara University was redesigned in the spring semester of 2024–2025 with a cohort of 11 final-year students to incorporate AI-supported semantic analysis. This initiative represents the first systematic integration of AI for semantic analysis within KFL curricula in Türkiye.

The aim of this study is to examine this pedagogical integration across three dimensions: (1) implementation feasibility: how AI tools were embedded into classroom practice; (2) student outcomes: linguistic development and motivational effects; (3) instructor perspective: pedagogical affordances and challenges encountered.

The significance of the study lies in its replicability as a model for KFL education in Türkiye and its transferability to other less commonly taught languages. By situating AI-supported semantic analysis within a real classroom context, this investigation offers empirical, evidence-based insights into both the opportunities and the limitations of employing AI in foreign language pedagogy.

2. Literature Review

The integration of artificial intelligence into language education has experienced substantial growth since 2021, with accelerated momentum following the release of ChatGPT in late 2022. This review examines studies published between 2021 and 2025, focusing primarily on English as a Foreign Language (EFL) and Korean as a Foreign Language (KFL) education. The literature clusters around three principal domains: AI-enhanced writing, technology acceptance and implementation, and affective outcomes.

Research on AI-enhanced writing demonstrates consistent improvements when combined with pedagogical scaffolding. Studies have documented effectiveness in grammar correction, content development, and revision processes (Han & Li, 2024; Hong & Shin, 2025; Kim, K., 2024; Wang, 2024; 김명희, 2023; 김성조, 2024). The "AI + Teacher" hybrid model emerges as particularly effective, with structured guidance producing superior outcomes compared to unrestricted access (Hong & Shin, 2025; Lin & Hwang, 2025). However, these studies focus predominantly on monolingual writing tasks rather than cross-linguistic semantic analysis required in translation pedagogy. Teacher and student perceptions reveal cautiously optimistic attitudes tempered by implementation concerns and ethical risks (Lee, 2024; Hınız, 2024; 류승의 et al., 2025). Experimental studies document positive affective outcomes, including reduced anxiety and enhanced critical thinking (Yu & Tao, 2025; Liu & Wang, 2024). Notably, KFL-focused studies were conducted primarily in Korea, China, and the United States, with learner populations predominantly East Asian or English-speaking (류승의 et al., 2025; 김보현, 2024).

Despite this growing body of research, none of the reviewed studies address AI-supported semantic analysis as an educational tool in translation training, particularly in KFL contexts where learners have no background in Chinese characters (Hanja). No empirical study has examined Turkish learners of Korean or documented how learners' engagement with AI for semantic tasks evolves over sustained practice. This study fills that gap by offering the first systematic investigation of AI-supported semantic analysis applied to translation training in a Turkish university KFL course, providing a replicable pedagogical model for under-resourced contexts.

3. Review of Curricula in Turkish Universities

As of 2025, the Department of Korean Language and Literature in Türkiye exists in three universities. The field of Korean language education as a foreign language and Korean Studies in Türkiye began in 1989 with the establishment of the Department of Korean Language and Literature at Ankara University, followed by Erciyes University, which commenced student admissions in 2003. Most recently, Istanbul University, founded in 2016, brought the field to its current form.

Although the teaching of Korean as a foreign language in Türkiye can be regarded as relatively well-positioned in a global context, the reality is that, in general, academic studies have not advanced much beyond a foundational level. For this reason, this study, though limited in scope, will present a general overview of the teaching of Korean as a foreign language in Türkiye under this heading.

First and foremost, it can be observed that intensive teaching of Korean as a foreign language

is carried out at all three universities. In addition, all three universities are also engaged in research in the field of Korean Studies. A close examination of the curricula reveals differences in distribution among the three universities. In recent years, Ankara University has stood out, particularly with the 2025 revision of its curriculum, by incorporating courses in Korean for Academic Purposes and Business Korean, thereby expanding into purpose-specific Korean language instruction. While Business Korean or similar courses are also offered at Erciyes and Istanbul Universities, Korean for Academic Purposes is absent. Similarly, a Korean course with a focus on artificial intelligence is not yet offered in the curricula of these three universities. However, the course “Comparative Semantic Studies” is currently being updated as a pilot course and is planned to be included in the course catalogue of Ankara University from 2026 onwards as an AI-based subject.

4. AI Integration in the Course Comparative Semantic Research (KRD422)

4.1 Course Syllabus

The Spring 2025 semester syllabus combined semantic theory with translation practice, supported by AI-enhanced tasks.

Table 1: Syllabus of the Course

Week	Date	Topic
1	17.02.2025	Introduction to the course content
2	24.02.2025	Semantic Concepts I (Denotation & Connotation, Synonymy & Antonymy)
3	03.03.2025	Semantic Concepts II (Polysemy, Homonymy, Meaning Change)
4	10.03.2025	Lexical Semantics (Word Relations, Semantic Components, Contextual Meaning)
5	17.03.2025	Translation Practice I (Brainwave Technology text)
6	24.03.2025	Translation Practice II (A Delayed Life text)
7	31.03.2025	Ramadan Holiday
8	07.04.2025	Midterm Exam
9	14.04.2025	Translation Practice III (Paternity leave text)
10	21.04.2025	Translation Practice IV (Emotional expressions text)
11	28.04.2025	Translation Practice V (Self-acceptance text)
12	05.05.2025	Translation Practice VI (Kaan Fighter jet text)
13	12.05.2025	Translation Practice VII (Constellations & Turkish Cuisine)
14	19.05.2025	National Holiday (Commemoration of Atatürk, Youth and Sports Day)
15	26.05.2025	Translation Practice VIII (Korean Geography & K-Beauty)

4.2 AI Prompts Integrated into the Course

Two primary prompt types scaffolded student learning throughout translation practice (Weeks 5–15):

Type 1: Comparative Evaluation Prompt

Students submitted their draft translations alongside the instructor's reference version, asking AI to evaluate across five criteria: stylistic appropriateness, grammatical accuracy, structural coherence, semantic similarity, and translation accuracy (each scored 0–20 points). Students also requested identification of spelling/orthographic errors.

Example: "Compare my translation with the reference version. Evaluate according to the five criteria, provide scores, and list any spelling errors in my Korean text."

Type 2: Semantic Analysis Prompts

Students used AI to explore lexical relationships and meaning structures:

- Semantic mapping: Identify vocabulary and map relationships (synonymy, antonymy, hypernymy)
- Field categorization: Classify words by semantic domain (technology, politics, culture)
- Contextual disambiguation: Clarify figurative language and grammatical structures

Example: "Identify key words in this Korean text and create a semantic map showing their relationships. Provide Turkish translations."

These prompts addressed common challenges: limited metalinguistic vocabulary, over-reliance on dictionary definitions, and lack of iterative feedback. Early weeks used pre-written templates; later weeks encouraged students to formulate their own queries based on specific translation difficulties.

4.3 Semantic Challenge Cases: Text Selection Process

The translation materials used in this course were drawn from diverse authentic sources, including Korean language teaching resources for foreign learners, online blogs, and news websites. This eclectic selection reflects the instructor's pedagogical aim of exposing students to varied registers, genres, and semantic challenges. Ten translation cases were integrated into the syllabus, each chosen to illustrate a distinct domain of language use.

Collectively, these cases encompassed technical/medical discourse (Brainwave Technology), sociocultural commentary (A Delayed Life, Paternity Leave), semantic-pragmatic analysis (Emotional Expressions, Self-Acceptance), current affairs and international cooperation (KAAN Fighter Jet), cultural knowledge (Constellations, K-Beauty), cross-cultural representation of Turkish society in Korean (Turkish Cuisine), and geographical description of Korea (Korean Geography). This diversity ensured students engaged with multiple registers, genres, and semantic challenges representative of Korean Studies scholarship.

Seven cases are presented below, representing the thematic categories outlined above.

Each case includes one representative excerpt in the original language (Korean or Turkish), its English translation, and ten key vocabulary items.

(1) Brainwave Technology (KO → TR)
Original (KO): 뇌파 인식 기술은 의료 분야에서 인간의 신체 기능을 대신하는 기기를 만드는 데 활용될 수 있다.
English: “Brainwave recognition technology can be applied in the medical field to create devices that substitute for human bodily functions.”
Vocabulary: 뇌파 (brain waves), 인식 (recognition), 기술 (technology), 의료 (medical), 분야 (field), 신체 (body), 기능 (function), 대표적 (representative), 장치 (device), 환자 (patient).

(2) Paternity Leave (KO → TR)
Original (KO): 최근 사표를 내는 것보다 더 어렵다는 '육아 휴직서'를 내는 아빠들이 늘고 있다.
English: “Recently, more fathers have been filing childcare leave applications, which are said to be even more difficult to submit than resignation letters.”
Vocabulary: 사표 (resignation letter), 어렵다 (difficult), 육아 (childcare), 휴직서 (leave application), 아빠 (father), 늘다 (to increase), 난생 (first time in life), 처음 (first time), 도전 (challenge), 직장인 (office worker).

(3) Self-Acceptance (KO → TR)
Original (KO): 다른 사람의 평가에 익숙해지는 것은 많은 사람들이 겪는 어려움 중 하나입니다.
English: “Becoming accustomed to others’ evaluations is one of the difficulties many people experience.”
Vocabulary: 사람 (people), 평가 (evaluation), 익숙해지다 (to get used to), 어려움 (difficulty), 겪다 (to experience), 자기 (self), 수용 (acceptance), 받아들이다 (to accept), 개인 (individual), 포용하다 (to embrace).

(4) KAAN Fighter Jet (KO → TR)
Original (KO): 인도네시아 아나타라 통신은 프라보워 대통령이 기자회견에서 "인도네시아는 튀르키예 방산업계와 함께 5세대 전투기 '칸' 개발과 잠수함 개발에 참여하기를 원한다"고 보도했다.
English: “Indonesia's Antara News Agency reported that President Prabowo, in a press conference, expressed his country's wish to participate in the development of the fifth-generation fighter jet 'KAAN' and submarine projects alongside Türkiye's defense industry.”
Vocabulary: 인도네시아 (Indonesia), 통신 (news agency), 대통령 (president), 기자회견 (press conference), 튀르키예 (Türkiye), 방산업계 (defense industry), 전투기 (fighter jet), 개발 (development), 잠수함 (submarine), 참여하다 (to participate).

(5) Constellations (KO → TR)
Original (KO): 별은 인류 역사 속에서 언제나 신비롭고 매혹적인 존재였습니다.
English: “Stars have always been mysterious and fascinating throughout human history.”
Vocabulary: 별 (star), 인류 (humanity), 역사 (history), 신비롭다 (mysterious), 매혹적 (fascinating), 밤하늘 (night sky), 수놓다 (to embroider), 패턴 (pattern), 별자리 (constellation), 부르다 (to call).

(6) Turkish Cuisine (TR → KO)
Original (TR): Türk toplumunda günlük yaşam, öğünlerin ve uzun yemeklerin etrafında şekillenir.
English: “In Turkish society, daily life revolves around meals and long dining occasions.”
Vocabulary: toplum (society), günlük (daily), yaşam (life), öğün (meal), şekillenmek (to be shaped), ev (home), hayat (life), mutfak (kitchen), aile (family), tercih etmek (to prefer).

(7) Korean Geography (KO → TR)
Original (KO): 한반도는 동북아시아 중심에서 서쪽의 중국, 동쪽의 일본 사이에 위치하며, 북위 33-43도, 동경 124-132도에 있다.
English: “The Korean Peninsula is located in Northeast Asia, between China to the west and Japan to the east, spanning 33–43 degrees north latitude and 124–132 degrees east longitude.”
Vocabulary: 한반도 (Korean Peninsula), 동북아시아 (Northeast Asia), 중심 (center), 중국 (China), 일본 (Japan), 위치하다 (to be located), 북위 (north latitude), 동경 (east longitude), 길이 (length), 면적 (area).

4.4 Student Learning Outcomes

At the end of the spring 2025 semester, as part of the final exam, students were given a 10-point question designed to encourage reflection on their learning experience. They were informed from the outset of the exam that the 10 points would be awarded automatically, and in return, they were asked to share their thoughts about the course directly and without any filtering. In other words, the aim was to access students’ perceptions of the course in their raw form. A total of 11 final-year undergraduate students, three male and eight female, provided brief written reflections, each ranging from 2 to 8 sentences.

These reflections, collected as part of routine course evaluation rather than formal research, were analyzed thematically to identify recurring patterns. To protect privacy, all names have been replaced with participant codes (S1–S11). While not designed as systematic research

data, these authentic student voices offer valuable insights into the pedagogical impact of AI-supported semantic analysis.

The instructor conducted inductive thematic analysis, using AI to assist with initial keyword extraction and pattern identification, followed by manual refinement and theme validation through multiple readings of student responses.
Thematic analysis identified nine recurring themes, summarized in Table 2.

Theme	n=11
Translation skills development	8
AI tool adoption	6
Vocabulary expansion	5
Future applicability	4
Academic language awareness	3
Course satisfaction	3
Motivational change	2
Self-awareness of gaps	2
L1 (Turkish) awareness	2

Table 2: Thematic Analysis of Student Reflections

Given the small sample size, these frequencies should be interpreted as illustrative rather than generalizable.

Sentiment analysis indicated predominantly positive perceptions (10 out of 11 students), with one student acknowledging challenges while also expressing determination to improve. Given the unbalanced gender distribution (3 male, 8 female), gender-based patterns are noted descriptively rather than comparatively. Representative student voices by theme included:

- On translation transformation: "Before this course, I would do word-for-word translation and move on. I realized how detailed a translation can be and how much each detail can change meaning." (S10)
- On AI as analytical tool: "I acquired the method of having my own translations analyzed and addressing my mistakes and deficiencies." (S9)
- On motivational shift: "My interest in Korean declined in previous years, but with the translations we did in class, it rose again… I caught my motivation." (S10)
- On L1 awareness: "I learned that translation requires not just knowing the subject's terminology very well, but also having a strong command of Turkish." (S7)
- On self-awareness and growth: "I struggled a lot because I have many deficiencies… but this course allowed me to practice. Even though I still have deficiencies, from now on I want to work on them and improve." (S4)
- On academic belonging: "For four years, only in your classes have I felt inside academic life. There was very high-quality content." (S5)

4.5 Instructor Perspective

Integrating artificial intelligence into the course initially presented notable pedagogical challenges for the instructor. Because the course content was centered on comparative semantics, it demanded substantial enhancement of students' comprehension of Korean as well as expansion of their Korean vocabulary. To support this development, the instructor compiled Korean–Turkish and Turkish–Korean translation texts, described in detail in the preceding section. Prior to these activities, theoretical lectures on the fundamentals of semantics were delivered over three weeks to establish a conceptual framework that would enable students to understand the logic of the course before engaging in practical applications.

Following the theoretical sessions, the course advanced through continuous analyses of semantic issues, during which AI was actively integrated. Students were engaged not only in individual written translations but also in collaborative work with peers of their choice. In this respect, the course served as a multi-dimensional pedagogical space for observation and experimentation.

The primary pedagogical challenge was students' limited lexical repertoire in both languages. To address this issue, the course adopted an iterative workflow:

1. Before class, the instructor selected the source text to be used.
2. With AI support, the instructor generated lists of less familiar vocabulary items (e.g., by prompting: "Identify TOPIK level 3 and above vocabulary and provide equivalents").
3. During class, the instructor presented the source text for translation.
4. Students began working on translations using traditional collaborative methods.
5. At points where students encountered significant difficulty, the instructor selectively provided the unknown-word list.
6. Translations were carried out in class with the support of vocabulary assistance and peer feedback.
7. Completed translations were evaluated through AI-based comparative feedback.

At this stage, the instructor also presented a reference translation refined with AI support and instructor review.

8. Finally, students reflected on the feedback received, engaging in discussion about AI use, translation monitoring, and strategies for more effective language use. Each class session concluded with this reflective activity.

AI played a dual role throughout the course: for students, it functioned as a rapid feedback provider (though sometimes overly technical in tone), and for the instructor, it served as a lexical assistant for vocabulary preparation and draft generation. This dual function created a balance between efficiency, since AI automation freed time for semantic analysis, and quality, as the instructor's expertise ensured accuracy and pedagogical appropriateness.

The workflow required ongoing calibration. At the beginning of the semester, students tended to rely excessively on AI output as authoritative. By mid-semester, however, most had developed a more critical stance, using AI as a point of departure for reflection rather than as a final answer, while continuing to benefit from it for self-development. Furthermore, contrary to public perception that younger generations use AI widely and with ease, the instructor observed that students' actual skills in employing AI were relatively limited and required substantial guidance to be effective.

5. Conclusion

This study examined the integration of AI-supported semantic analysis into Korean as a Foreign Language (KFL) education in Türkiye through the redesign of the Comparative Semantic Research course (KRD422) at Ankara University. By situating AI tools within a semester-long translation-focused curriculum, the study addressed a critical gap in the literature: while existing research on AI in language education has predominantly concentrated on grammar correction, writing, or conversational practice, semantic competence, particularly in translation and semantic analysis, remained underexplored.

Student reflections revealed that AI integration enhanced translation competence, expanded lexical repertoires, and fostered metalinguistic awareness. Thematic analysis identified improvements across nine dimensions, with translation skills development and AI tool adoption emerging as dominant themes. Students reported gains not only in technical proficiency but also in motivation, academic belonging, and self-awareness of learning gaps. From the instructor's perspective, AI functioned as a dual resource for lexical preparation and comparative feedback. Critically, students evolved from over-reliance on AI outputs to more reflective use, indicating that guided integration can cultivate both linguistic and metacognitive growth.

The contribution of this study is twofold. First, within the Turkish context, it represents the first systematic attempt to incorporate AI-supported semantic analysis into a KFL curriculum, offering a replicable model for other institutions. Second, within the broader field of language education, it demonstrates how AI can be embedded into meaning-focused pedagogy, extending current applications beyond grammar or fluency to the level of semantic depth and intercultural interpretation.

Several limitations must be acknowledged. The study was exploratory in design, based on a single institution, one elective course, and one semester, which restricts the generalizability of its findings. Data were collected as part of routine course evaluation rather than through systematically designed research instruments. Future research should broaden the scope of AI integration to multiple KFL courses and across multiple universities. Longitudinal and comparative designs contrasting AI-supported and traditional methods would provide stronger evidence of pedagogical effectiveness.

In conclusion, while focused in scope, this study provides a reference point for how AI can be employed to enhance semantic competence and translation awareness in KFL education. Beyond the Turkish–Korean context, the pedagogical framework demonstrated here holds potential for adaptation to other less commonly taught language pairs, particularly where resources for specialized semantic instruction are limited.

References

Han, J., & Li M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. System, 126. https://doi.org/10.1016/j.system.2024.103502
Hınız, G. (2024). A year of generative AI in English language teaching and learning - A case study. Journal of Research on Technology in Education. https://doi.org/10.1080/15391523.2024.2404132
Hong, S., & Shin, Y. K. (2025). Effects of three levels of AI integration on second language academic writing: Evaluating restricted, guided, and free use of ChatGPT. System, 134. https://doi.org/10.1016/j.system.2025.103820
Kim, K. (2024). The potential of generative AI in writing feedback for Korean L2 learners: An analysis on grammar error correction by ChatGPT-3.5 for TOPIK II writing tasks. In Innovative Methods in Korean Language Teaching (pp. 64–76). Taylor and Francis. https://doi.org/10.4324/9781032725307-6
Lee, I. (2024). AI-powered writing assistance: Korean language students' and teachers' views and experiences. In Innovative Methods in Korean Language Teaching (pp. 13–27). Taylor and Francis. https://doi.org/10.4324/9781032725307-3
Lin, C.-J., & Hwang, G.-J. (2025). Artificial intelligence-supported procedural scaffolding for promoting EFL learners' writing performance in flipped peer assessment activities. Interactive Learning Environments, 1–15. https://doi.org/10.1080/10494820.2025.2532629
Liu, W., & Wang, Y. (2024). The Effects of Using AI Tools on Critical Thinking in English Literature Classes Among EFL Learners: An Intervention Study. European Journal of Education, 59(4). https://doi.org/10.1111/ejed.12804
Wang, D. (2024). Teacher-Versus AI-Generated (Poe Application) Corrective Feedback and Language Learners' Writing Anxiety, Complexity, Fluency, and Accuracy. In International Review of Research in Open and Distributed Learning (Vol. 25).
Yu, J., & Tao, Y. (2025). To be in AI-integrated language classes or not to be Academic emotion regulation. British Educational Research Journal, 00, 1–26.
김명희. (2023). ChatGPT를 활용한 한국어 글쓰기 교수-학습 방안 연구. 한국문예창작, 22(2), 55–86.
김보현. (2024). 생성형 AI 기반 에듀테크의 한국어교육 적용 가능성과 쟁점 -대학 한국어 교사의 인식과 생성형 AI 의 특성을 중심으로. 교양 교육 연구, 18(6), 291–307. https://doi.org/10.46392/kjge.2024.18.6.291
김성조. (2024). 인공지능(AI)을 활용한 한국어교육 방안 연구. 외국어로서의 한국어교육, 75, 33–58. https://doi.org/10.21716/tkfl.75.2
류승의, 박아희, & 조윤경. (2025). 한국어 교육에서의 AI 활용에 대한 한국어 교사들의 인식 연구. 언어와 문화, 21(3), 117–145. https://doi.org/10.18842/kl

AI에 대한 그리스도교적 성찰

A Christian Reflection on Artificial Intelligence

방종우
가톨릭대학교 교수

Jongwoo Bang
Professor, The Catholic University of Korea Songsin Campus

초록

가톨릭교회는 AI 발전에 대응하여 『로마의 호소』와 「옛것과 새것」을 통해 윤리적 성찰을 제시하며, AI의 책임이 개발자, 규제 기관, 사용자 의식에 달려 있음을 강조한다. 핵심은 AI가 인간 지능을 대체할 수 없으며, 인간의 창조성과 도덕적 책임을 고양시키는 도구가 되어야 한다는 점이다. 특별히 교회는 '지능' 용어의 혼동에 경고하며 인간과 AI 지능의 본질적 차이를 명확히 한다. AI 지능은 데이터 기반의 기능적 수준이며, 인간 정신을 환원하는 기능주의적 위험을 내포한다. 한편 인간 지능은 이성(ratio)과 지성(intellectus)을 통해 진리를 탐구하고 윤리적 판단을 내리는 총체적 능력이자 "진리의 터득을 위한 고유한 선물"이다. 한편 AI는 추론을 모방하지만, 도덕적 분별력, 공감, 통합적 이해 등 인간 마음의 풍요로움이 결여되어 있다. 따라서 AI는 인간 지능의 '결과물'로 간주되어야 한다. 교회는 진정한 지혜를 추구해야 한다고 권고하며, 지혜는 정보의 양이 아닌 "그가 행하는 사랑의 깊이"에서 나오며, 연대와 공동 이익 증진을 통해 구현된다는 사실을 강조한다. 우리가 기술을 개발하고 활용하는데 윤리를 어떻게 구체적으로 개발하고 적용하는 가에 따라, AI가 희망찬 지혜의 원동력이 될지, 치명적인 실패가 될지가 판가름 날 것이다.

Abstract

In response to the rapid advancement of artificial intelligence (AI), the Catholic Church, through the Rome Call for AI Ethics and Antiqua et Nova, offers ethical reflections emphasizing that responsibility for AI lies with developers, regulatory bodies, and users' moral awareness. At the core is the conviction that AI can never replace human intelligence; rather, it should serve as a tool to elevate human creativity and moral responsibility. The Church particularly warns against the semantic confusion surrounding the term intelligence and insists on clarifying the essential distinction between human and artificial intelligence. AI operates at a functional, data-driven level and carries the risk of functionalism that reduces the human mind to mere computation. Human intelligence, by contrast, encompasses ratio and intellectus, enabling the pursuit of truth and the exercise of ethical judgment as a "unique gift for the attainment of truth." While AI can imitate reasoning, it lacks the moral discernment, empathy, and integrative understanding that characterize the richness of the human spirit. Therefore, AI should be regarded as a product of human intelligence rather than its equivalent. The Church calls for the pursuit of true wisdom, which is not measured by the amount of information but emerges from "the depth of love expressed in action," manifested through solidarity and the promotion of the common good. Ultimately, the development and application of ethical frameworks will determine whether AI becomes a source of hope-filled wisdom or leads to catastrophic failure.

1. 가톨릭교회의 움직임

인간의 삶에 지대한 영향을 미치는 과학 기술의 발전 속에서 교회는 사회 회칙을 비롯한 다양한 기회를 통해 인간의 존엄성과 생명의 가치에 윤곽을 제시해 왔다. 특별히 AI와 관련하여 교황청 생명 아카데미는 2020년, AI 기술의 시의성에 기대와 우려를 표명하며 해당 분야의 몇몇 기업과 단체들과 함께[1] 「AI 윤리에 대한 로마의 호소」(Rome Call for AI Ethics, 이하 「로마의 호소」)[2]를 발표하였다. 이후, 약 5년이라는 시간이 흘렀고 그동안 과학기술은 또 다른 새로운 기능들을 탄생시킴으로써 선형적인 속도가 아닌 기하급수적 속도로 진화되었다. 이러한 현상은 인류에게 새로운 문제와 의문들을 가져왔다. 이에 따라 AI 윤리에 관한 사회적 논의가 지속되었으며 그에 따라 각 기업 혹은 국가는 저마다의 원칙을 세우고자 했다. 하지만 그럼에도 불구하고 여전히 AI의 폐해가 우려되며 윤리적 통제가 가능한가에 대한 문제 역시 제기되고 있는 실정이다. 기본적으로 AI는 데이터의 축적과 알고리즘을 기반으로 하고 있으며 이는 한 기업에 속한 개별 인간 개발자들에 의해 설계되기 때문이다. 또한 데이터의 축적으로 개발자의 의향과 상관없이 예측할 수 없는 방향으로 나아가기 때문이다. 결국 AI 기술의 윤리적 문제들과 이를 최소화하기 위한 방편은 개발자의 자발적 의지, 상위 기관의 규제 의지, 사용자의 의식과 식별 능력에 달려있다고 할 수 있다.

이러한 시대의 필요에 입각해 교황청 신앙교리부와 문화교육부는 최근 2025년 1월 14일, 인간의 지능과 인공지능의 관계에 대한 가이드라인, 「옛것과 새것」(*Antiqua et Nova*)[3]을 발표했다. 이 문헌은 AI의 기술적 발전 자체를 단순히 환영하거나 비판하는 차원을 넘어, 인간의 지능과 자유, 그리고 하느님의 모상(Imago Dei)이라는 인간 이해에 비추어 AI가 제기하는 도전과 기회를 인간학적으로 재해석한다. 결국 이 문헌은 AI가 인간의 지능을 대체하는 것이 아니라, 인간의 창조성과 도덕적 책임을 더욱 고양시키는 방향으로 발전해야 한다는 점을 강조하며, 신앙과 이성, 전통과 혁신이 조화롭게 만나는 새로운 윤리적 지평을 열고자 한다. 이에 본 발표문에서는 가톨릭교회가 이야기하는 '지능'에 관한 부분에 초점을 맞추어 그에 대한 성찰을 시도해보고자 한다.

2. 인간의 지능, 기계의 지능

모두가 알고 있듯 AI(인공지능)는 '지능'이라는 용어를 기계에도 역시 적용한 것이다. 이에 교회는, 인간의 지능과 기계의 지능을 동등한 위치에 둠으로써 오히려 숙고해야 할 윤리적 관점들이 제외될 수 있음을 지적한다. 즉, 인간의 지능은 그 사람의 전체와 관련된 능력인 반면, AI의 지능은 기능적인 수준에 머무르는데, 용어의 혼용으로 인간 정신의 고유한 활동까지 기계가 수행할 수 있다고 오해할 위험이 있는 것이다.[4] 좀 더 구체적으로, 이러한 접근 방식은 인간의 마음을 기능으로 환언하고 이를 물리적, 수학적으로 정량화할 수 있다고 가정하는 기능주의적 관점을 반영한다.[5]

---

1) 이 문서가 발표되던 당시, 빈센조 팔리아(Vincenzo Paglia) 교황청 생명 아카데미 의장, 브래드 스미스(Brad Smith) 마이크로소프트 최고법률책임자(CLO), 존 켈리(John Kelly) IBM 수석 부사장, 쿠 동규(Qu Dongyu) FAO 사무총장, 파올라 피사노(Paolo Pisano) 이탈리아 혁신부 장관 등이 함께 서명하였다.

2) Pontifical Academy for Life, "Rome Call for AI Ethics", (28 febbraio 2020), https://www.romecall.org/wp-content/uploads/2021/02/AI-Rome-Call-x-firma_DEF_DEF_con-firme_.pdf[2021.10.18.].

3) Dicastero per la dottrina della fede · Dicastero per la cultura e l'educazione, Antiqua et nova, (18 gennaio 2025), https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_it.html[2025.2.13.].

4) Cf. Antiqua et Nova, n.10.

5) 기능주의(functionalism)는 정신 상태를 그것이 수행하는 기능적 역할, 즉 감각 자극에 대한 반응, 다른 정신 상태와의 관계, 그리고 행동적 산출물과의 관계를 통해 정의하는 이론이다. 이 관점에 따르면, 정신은 특정 물질적 구성에 귀속되지 않으며, 동일한 기능적 구조를 구현할 수 있다면 어떤 물리적 시스템에서도 정신 상태가 발생할 수 있다. 이러한 입장은 AI가 인간과 유사한 정신 능력을 가질 수 있다는 가능성을 이론적으로 정당화하는 데 자주 인용된다. 그러나 기능주의는 인간 정신을 특정 물질적 기반이 아니라 기능적 구조와 역할에 따라 정의한다는 점에서, 인간의 정신적·도덕적 실재를 축소시키는 위험을 내포한다. 인간의 정신은 단순한 정보 처리 기계적 기능이 아니라, 인격적 자아, 자유의지 안에서 이해되어야 하기 때문이다. 참조: 교황청 AI 연구 그룹, 『인공지능과 만남』, 이성효 외 9인 옮김, 수원: 수원가톨릭대학교 출판부, 2025, 100-101쪽.

그러므로 AI가 많은 능력을 보유하고 있으며 나아가 미래의 일반인공지능(AGI, Artificial General Intelligence)이 인간과 유사한 모습을 보일 수 있다고 할지라도, 본질적으로는 이 기술이 여전히 기능적이라는 점에 주목할 필요가 있다. 비록 AI가 대규모 데이터를 분석하고 패턴을 파악함으로써 그 결과를 예측하고 새로운 접근 방식을 제안함으로써 인간의 문제 해결 능력의 전형적인 특정 인지 과정을 모방할 수 있다고 하더라도, 공감하는 능력, 양심의 소리를 따르는 능력은 없는 것이 사실이다. 그러므로 인간의 지능과 기계의 지능은 결정적인 차이가 존재함을 염두에 두어야 한다.

고전적 전통에서 지능(intelligence)의 개념은 단순한 정보 처리 능력이 아니라 진리를 탐구하고 윤리적 판단을 내리는 사고 능력, 즉 흔히 '이성'(ratio)과 '지성'(intellectus)이라는 상보적인 개념을 통해 이해된다. 플라톤(Platon)은 '지성에 의한 이해(앎, 직관)'(noesis)가 진리를 파악하는 가장 고차원적인 능력이며 감각적 경험이 아닌 순수한 이성을 통해 작용하는 것으로, 철학적 탐구를 통해 계발된다고 설명한다.[6] 한편 아리스토텔레스(Aristotle)는, 영혼 안에서 행위와 진리를 지배하는 세 가지[7] 중 하나를 지성(Nous, νοῦς)이라고 부르며, 특별히 지적 활동과 관련하여 '실천적 지혜'(Phronesis, φρόνησις)와 '학문적 인식'(Episteme, ἐπιστήμη)을 언급한다. 여기서 실천적 지혜란, 윤리적 판단과 관련된 신중한 사고 능력이며 윤리덕을 바탕으로 올바른 행동을 결정하는 능력이다. 한편 학문적 인식이란 순수한 논리적·철학적 탐구 능력을 의미한다. 특별히 아리스토텔레스는 지성적 덕(intellectual virtues)을 강조하며, 인간의 지능이 경험을 통해 발전한다고 여겼다.[8] 토마스 아퀴나스(Thomas Aquinas)는 지혜를 "이성"과 "지성"이라는 상호보완적인 개념으로 표현한다. 그에 따르면 이성과 지성은 별개의 분리된 능력이 아니라, 지혜가 작용하는 두 가지 방식이다. "지성과 이성은 비록 서로 다른 능력들이 아니지만, 그럼에도 불구하고 서로 다른 행위들에 근거해서 명명되었다. 지성이라는 이름은 '진리의 가장 깊숙한 통찰'에서 취해졌고, 반면 이성이라는 이름은 탐색과 논변에서 왔다."[9]

이를 종합해 봤을 때, 전통적인 철학 개념 안에서 지능(지적 이해 능력)이란, 단순히 인지능력, 논리적 사고, 문제 해결 능력, 학습 능력, 기억력이나 정보 처리 능력뿐만이 아닌, 본질을 이해하는 능력, 직관적 통찰, 사유 능력, 윤리적 사고와 논리적 탐구, 그리고 새로운 개념을 형성하는 능력을 모두 포괄하고 있음을 알 수 있다. 이는 곧 인간의 지적 이해 능력이 인간 활동의 모든 측면을 형성하고 관통한다는 것을 의미한다. "지성은 성찰과 경험과 대화를 통하여 사물들의 실재를 통찰할 수 있고, 그 실재 안에는 이를 초월하는 어떤 보편적 도덕적 요구들의 기초가 있음을 인식할 수 있다."[10] 이 포괄적인 관점은 인간이 어떻게 자신의 의지와 행동을 향상시키고, 형성하고, 변화시키는 방식으로 통합되는지를 보여준다.[11]

이러한 관점에서 교회의 가르침을 보면, 인간의 지능은 궁극적으로 "진리의 터득을 위하여 만들어진 하느님의 선물"[12]이다. 즉, 인간의 지능이 지닌 이중 구조인 지성과 이성(intellectus-ratio)은 인간이 단순한 감각적 경험이나 효용성을 넘어서는 진리를 탐구할 수 있게 한다. 왜냐하면 진리에 대한 갈망은 인간 본성의 일부이기 때문이다. "사물들이 왜 지금 있는 모습으로 있는 것인지를 묻는 것은 인간 이성의 타고난 속성이다."[13] 이렇게 진리를 추구하려는 이러한 인간 지능의 타고난 욕구는 특히 의미 이해와 창의성의 특유한 인간적 능력에서 두드러지며, 이는 인간 본성의 사회적 성격과 존엄성에 부합하는 방식으로 전개된다.[14]

결론적으로 인간의 지능은 단순한 사실 습득 능력이나 특정한 과제를 수행하는 능력으로 축소될 수 없다. 인간은 삶의 궁극적인 질문에 열려 있으며, 참되고 선한 것을 향한 방향을 반영한다.[15] 인간의 지능은 존재의 총체성에 접근할 수 있는 능력을 가지고 있으며, 측정 가능한 범위를 넘어 충만함 속에서 존재를 숙고하고 이해의 의미를 파악할 수 있다. 그리스도교인들에게 이 능력은 특별한 방식으로 이성을 사용해 계시된 진리에 더욱 깊이 관여하게 함으로써 하느님의 신비를 더욱 깊이 인식할 수 있는 능력(intellectus fidei, 신앙의 이해)으로 드러난다.[16] 이를 통해 인간 지능은 본질적으로 관조적인 차원을 지니고 있으며, 어떤 실용적인 목적보다 진리, 선, 그리고 아름다움에 대한 이타적인 개방성을 지니고 있다는 것을 알 수 있다.

3. AI의 한계

앞서 논의한 내용을 고려할 때, 인간 지능과 인공 지능의 차이가 분명해진다. AI는 인간 지능과 관련된 특정 결과를 모방할 수 있는 탁월한 기술적 성과이지만 정량적 데이터와 전산 논리를 기반으로 작업을 수행할 뿐이며, 이를 바탕으로 결정을 내리는 방식으로 작동한다. 결국 AI가 인간 지능의 일부를 처리하고 모방하며 설사 일반인공지능(AGI, Artificial General Intelligence)이 개발된다 할지라도, 근본적으로 논리-수학적 틀에 국한되어 있기 때문에 본질적인 한계가 있는 것이다. 반면에 인간의 지능은 신체적, 심리적 성장 과정 전반에 걸쳐 유기적으로 발달하며, 수많은 삶의 경험을 통해 형성된다. 고도로 발전한 AI 시스템은 머신러닝(Machine Learning)과 같은 과정을 통해 학습을 할 수 있지만, 이러한 훈련은 감각적 경험, 정서적 반응, 사회적 교류, 각 순간의 고유한 맥락 등 구체화된 경험을 통해 형성되는 인간 지능의 발달적 성장과는 근본적으로 다르다.[17]

결과적으로, AI는 인간의 추론 능력을 모방하고 놀라운 속도와 효율성으로 특정 작업을 수행할 수 있지만, 그 연산 능력은 인간 정신의 광범위한 능력의 극히 일부에 불과하다. 예를 들어 AI는 도덕적 분별력을 수행할 수 없으며 진정한 관계도 구축할 수 없다. 한편 인간의 지능은 개인의 지적, 도덕적 관점을 근본적으로 형성하는 개인적 삶의 역사에 기반을 두고 있으며, 삶의 신체적, 정서적, 사회적, 도덕적, 영적 차원을 포괄한다. 이에 「옛것과 새것」은 다음과 같이 말한다.

---

6) 참조: 플라톤, 『국가』, 박종현 역주, 서울: 서광사, 2011, 제7권, 514a~524e.

7) 감각(aisthesis), 지성(nous), 욕구(orexis)

8) 참조: 아리스토텔레스, 『니코마코스 윤리학』, 최명관 옮김, 서울: 창, 2008, 1139b.15-1140b.25.

9) 토마스 아퀴나스, 『신학대전』, 이상섭 옮김, 횡성: 한국성토마스연구소, 2023, II-II, q. 49, a.5, ad 3.

10) 프란치스코,「모든 형제들」, 서울: 한국천주교중앙협의회, 2020, 213항.

11) Cf. Congregazione per la Dottrina della Fede, "Nota dottrinale su alcuni aspetti dell'evangelizzazione", (3 dicembre 2007), n.4, https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_con_cfaith_doc_20071203_nota-evangelizzazione_it.html[2025.3.16.].

12) Congregazione per la Dottrina della Fede, Donum veritatis, (24 maggio 1990), n. 6, https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_con_cfaith_doc_19900524_theologian-vocation_it.html[2025.3.16.].

13) 요한바오로 2세, 「신앙과 이성」, 서울: 한국천주교중앙협의회, 1998, 3항.

14) 인간에게 있어 의미를 수용하는 능력은 의사소통으로 표현된 메시지의 내용을 정보적 기호와 같은 물질적 또는 경험적 구조를 연관하는 방식으로 파악하는 동시에 그것을 초월할 수 있게 한다. 이 경우, 지능은 하느님의 눈으로 사물을 보고, 연결고리, 상황, 사건을 이해하고, 그 의미를 발견할 수 있게 해주는 지혜가 된다. Cf. Francesco, "Messaggio per la LVIII Giornata Mondiale delle Comunicazioni Sociali", (24 gennaio 2024), https://www.vatican.va/content/francesco/it/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html[2025.3.16.].

15) 인간 존엄성은 "하느님께서 당신 지혜와 사랑의 계획으로 온 우주와 인간 사회에 질서를 세우시고 이를 이끄시고 다스리시는 영원하고 객관적이며 보편적인 하느님 법을 인간 생활의 최고 규범이라고 여기는 사람에게 더욱 분명하게 드러난다. 하느님께서는 인간이 당신의 인자하신 섭리로 불변의 진리를 더욱더 깊이 깨달을 수 있도록 당신의 이 법에 인간을 참여시키셨다." 제2차 바티칸 공의회, 「인간 존엄성」, 1965, 3항.

16) Commissione Teologia Internazionale, "La teologia oggi: Prospettive, Principi e criteri", (2011.11.29), n.17, https://www.vatican.va/roman_curia/congregations/cfaith/cti_documents/rc_cti_doc_20111129_teologia-oggi_it.html[2025.3.16.].

17) Cf. Antiqua et Nova, n.28.

AI는 물질성, 관계성, 그리고 진리와 선에 대한 인간 마음의 개방성에서 오는 풍요로움이 부족하기 때문에, 그 능력이 무한해 보일지라도 현실을 파악하는 인간의 이해 능력과 비교될 수 없다. 인간은 질병, 화해의 포옹, 심지어 단순한 일몰에서도 많은 것을 배울 수 있다. 실제로 우리가 인간으로서 겪는 수많은 경험은 새로운 지평을 열고 새로운 지혜를 얻을 수 있는 가능성을 제공한다. 데이터로만 작동하는 어떤 장치도 우리 삶에 존재하는 이러한 경험과 셀 수 없이 많은 다른 경험들을 따라올 수 없다.[18]

그러므로 특별히 교회는 AI가 이러한 통합적 이해를 제공할 수 없기 때문에, 이 기술에만 의존하거나 이 기술로 세상을 해석하는 접근 방식은 "전체에 대한 감각, 사물들의 관계에 대한 감각, 넓은 지평에 대한 감각을 잃어버리는"[19] 결과를 초래할 수 있음을 경고한다. 나아가 인간의 지능과 AI를 지나치게 동일시하는 것에 대해, 인간이 수행하는 작업에 따라 그 가치를 매기는 기능주의적 관점에 빠질 위험이 있음을 우려한다. 이러한 관점에서 AI를 인간의 지능을 인공적으로 만든 형태로 간주해서는 안 되며, 인간 지능의 '결과물'로서 간주해야 한다. 이는 인간의 권리에 대해 "공통 기반을 찾는 데 있어 중요한 접점"[20]을 나타내며, 책임 있는 AI 개발과 사용에 관한 논의에 있어 필요한 윤리적 지침의 기초 정신을 제공한다.

4. 진정한 지혜를 위하여

교회는 진리와 더욱더 멀어지는 인간 사회에 대한 우려를 표하면서, 인간이 진리를 인식하지 못한다면 "인간 자신이 어디로 가고 있는지 알지 못하며 심지어는 자신이 누구인지도 발견하지 못한다"[21]고 이야기한다. 그리하여 타인에 대한 개방적인 태도와, 인간 스스로 대한 경외심을 가질 것을 강조한다. 그렇게 될 때에 인간은 자신의 한계에 대한 유혹에 빠지지 않고, 자기 자신을 겸허히 받아들임과 동시에 인간을 진정으로 발전시키는 방향으로 창의성의 산물인 기술을 올바로 '책임 있게' 개량해 나갈 수 있을 것이기 때문이다.

한편 '인간 개인'은 현재에 일시적으로 존재한다고 할 수 있지만 '인류'는 미래에도 존속해야 할 피조물이다. 이에 인간은 연대성 안에서 자기 자신을 포함한 모든 피조물과의 총체성과 연속성으로 인해 모든 미래에 대한 책임이 있다. 특별히 과학 기술은 이러한 인간의 조건에 매우 큰 영향을 미치므로 현대의 인류는 이에 대해 더 많은 책임을 요구받는다고 할 수 있다. 이는 AI 기술이, 현재에 있는 혹은 미래에 있을 타인 혹은 다른 피조물을 훼손하지 않고 지키는 한에서 개발되고 사용되어야 함을 뜻한다. 그리고 현재 인류는 이 질서를 어떻게 마련할 것인가에 대한 중대한 기로 앞에 서있다.

한편 과학적 진보가 인간적, 정신적 불모의 땅으로 남지 않도록 하기 위해서는 단순한 데이터의 축적을 넘어 진정한 지혜를 추구하고자 노력해야 한다. 그러므로 프란치스코 교황은 다음과 같이 언급한다.

우리에게 밀어닥치는 정보의 홍수는 더 큰 지혜로 이어지지 않습니다. 지혜는 재빠른 인터넷 검색에서 태어나는 것도 사실 검증이 이루어지지 않은 많은 자료도 아닙니다. 그러한 방법으로는 진리와의 만남을 통한 성숙이 이루어지지 않습니다.[22]

결국 이 시대의 중대한 기로 앞에서 우리는 참다운 지혜는 무엇인지에 대한 의식을 끊임없이 가져야 할 필요가 있다. 지혜란 더 완전한 사람이 되는 것을 목표로 하지만 그 척도는 단순한 지식의 양을 의미하는 것이 아니다. "사람의 완벽함의 척도는 그가 습득한 정보나 지식의 양이 아닌 그가 행하는 사랑의 깊이이다".[23] 이렇듯 지혜는 공동의 이익을 증진하고, 공동의 집을 돌보고, 진리를 추구하며 인류의 통합적 발전을 지원하고, 연대와 인류애를 증진시킬 때 비로소 모습을 드러낸다. 그러므로 기술을 보다 인간 증진을 위한 올바른 도구로 사용할 때 우리는 비로소 진정한 지혜를 소유하게 될 것이다. 따라서 우리가 기술을 개발하고 활용하는데 윤리를 어떻게 구체적으로 개발하고 적용하는 가에 따라, AI가 희망찬 지혜의 원동력이 될지, 치명적인 실패가 될지가 판가름 날 것이다.

18) Antiqua et Nova, n.33.

19) 프란치스코, 「찬미 받으소서」, 서울: 한국천주교중앙협의회, 2015, 110항.

20) Francesco, Discorso ai partecipanti all'Assemblea Plenaria della Pontificia Accademia per la Vita, (28 febbraio 2020), https://www.vatican.va/content/francesco/it/speeches/2020/february/documents/papa-francesco_20200228_accademia-perlavita.html[2025.3.16.].

21) 교황 베네딕토 16세, 「진리 안의 사랑」, 서울: 한국천주교중앙협의회, 2009, 78항.

22) 프란치스코, 「모든 형제들」, 50항.

23) Antiqua et Nova, n.116.

참고문헌

교황청 AI 연구 그룹, 『인공지능과 만남』, 이성효 외 9인 옮김, 수원: 수원가톨릭대학교 출판부, 2025.
교황 베네딕토 16세, 「진리 안의 사랑」, 서울: 한국천주교중앙협의회, 2009.
아리스토텔레스, 『니코마코스 윤리학』, 최명관 옮김, 서울: 창, 2008.
요한바오로 2세, 「신앙과 이성」, 서울: 한국천주교중앙협의회, 1998.
제2차 바티칸 공의회, 「인간 존엄성」, 1965.
토마스 아퀴나스, 『신학대전』, 이상섭 옮김, 횡성: 한국성토마스연구소, 2023.
프란치스코, 「찬미 받으소서」, 서울: 한국천주교중앙협의회, 2015.
프란치스코, 「모든 형제들」, 서울: 한국천주교중앙협의회, 2020.
플라톤, 『국가』, 박종현 역주, 서울: 서광사, 2011.
Commissione Teologia Internazionale, “La teologia oggi: Prospettive, Principi e criteri”, (2011.11.29), https://www.vatican.va/roman_curia/congregations/cfaith/cti_documents/rc_cti_doc_20111129_teologia-oggi_it.html.
Congregazione per la Dottrina della Fede, Donum veritatis, (24 maggio 1990), https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_con_cfaith_doc_19900524_theologian-vocation_it.html.
Congregazione per la Dottrina della Fede, “Nota dottrinale su alcuni aspetti dell’evangelizzazione”, (3 dicembre 2007), https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_con_cfaith_doc_20071203_nota-evangelizzazione_it.html.
Dicastero per la dottrina della fede · Dicastero per la cultura e l’educazione, Antiqua et nova, (18 gennaio 2025), https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_it.html.
Francesco, Discorso ai partecipanti all’Assemblea Plenaria della Pontificia Accademia per la Vita, (28 febbraio 2020), https://www.vatican.va/content/francesco/it/speeches/2020/february/documents/papa-francesco_20200228_accademia-perlavita.html.
Francesco, “Messaggio per la LVIII Giornata Mondiale delle Comunicazioni Sociali”, (24 gennaio 2024), https://www.vatican.va/content/francesco/it/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html.
Pontifical Academy for Life, “Rome Call for AI Ethics”, (28 febbraio 2020), https://www.romecall.org/wp-content/uploads/2021/02/AI-Rome-Call-x-firma_DEF_DEF_con-firme_.pdf.

저자의 두 번째 죽음: 인문학 연구에서 인간과 합성 지능의 공존을 향하여

The Second Death of the Author: Toward Coexistence Between Human and Synthetic Intelligence in Humanities Research

발라가나파티 데바라콘다
델리대학교 교수

Balaganapathi Devarakonda
Professor, University of Delhi

Abstract

The rise of generative AI has initiated a profound epistemic transformation within the humanities, challenging traditional notions of authorship, interpretation, and originality. This paper examines the impact of this transformation through what it terms the second death of the author, a condition in which authorship itself becomes indiscernible amid algorithmic opacity. Tracing the conceptual evolution of the humanities from studia humanitatis to posthuman thought, the paper situates the current crisis within a long genealogy of knowledge, interpretation, and human self-understanding. It argues that generative AI threatens not merely to automate textual production but to dissolve the hermeneutic foundation of humanistic inquiry. In response, the paper proposes pathways toward coexistence: reconfiguring authorship as relational ecology, foregrounding reflexive praxis over textual output, and reaffirming the ethical and temporal dimensions of human understanding. The future of the humanities, it contends, lies not in resisting AI but in redefining interpretation as a shared, reflective act within a hybrid ecology of intelligence.

I. Introduction: The New Epistemic Shock

The conceptual foundations of metaphysics, epistemology, and axiology within the humanities are undergoing radical transformation with the rise of Artificial Intelligence, particularly generative AI. If globalization marked the first major disruption to the humanities by unsettling their geographical and cultural centers, AI constitutes the second wave by transformation of their ontological and epistemic core. Where globalization expanded the *where* of humanistic inquiry, AI challenges the who and how of its authorship. Generative AI's capacity to produce original content which renders syntactically precise, stylistically sophisticated, and contextually adaptable output has blurred the boundaries between human creativity and algorithmic generation. This transformation poses questions that cut to the marrow of humanistic research as to what becomes of interpretation when machines can simulate it? What happens to originality when textual production becomes automated? And most urgently, how can speculative, meaning-oriented disciplines coexist with machinic systems that imitate meaning without understanding it?

This paper investigates the crisis through the lens of what may be called the second death of the author. Roland Barthes' original provocation in 1967 announced the symbolic "death" of the author as the origin of meaning, freeing the text for multiple interpretations. In contrast, the generative turn in AI effects a literal and ontological dissolution of authorship: a condition where the author is not merely de-centered but rendered indiscernible. What emerges is not liberation but *absorption* of the disappearance of authorship into algorithmic opacity. The implications for humanities research are existential, especially for its paradigmatic institution, the doctoral degree, which historically certifies interpretive originality and intellectual agency.

Yet, this paper does not treat the rise of AI as a catastrophe but as a philosophical provocation. The humanities must confront, not evade, the challenge of synthetic cognition by rethinking their own epistemic methods and ethical responsibilities. The argument proceeds in four movements: first, an exposition of the problem of authorship and epistemic opacity; second, a conceptual history tracing the evolution of the humanities from *studia humanitatis* to posthuman thought; third, a theoretical analysis of the AI–humanities encounter; and finally, an exploration of possible paths to coexistence through reflexive, relational, and ethical reconfiguration.

II. The Problem: The Second Death of the Author

Humanistic research has historically grounded itself in what may be called hermeneutic agency which is the ability to interpret, contextualize, and synthesize meaning through situated consciousness. A doctoral dissertation exemplifies this agency as it represents not just the reproduction of knowledge but also the performance of thought and a demonstration of interpretive responsibility within a living tradition.

Generative AI, however, destabilizes this foundation because large language models can now produce essays that imitate disciplinary idioms, simulate argumentative structure, and even replicate styles of reasoning across philosophy, literature, or cultural theory. The distinction between human-authored and AI-authored research is no longer determined by form or fluency but rather by the *opacity* of the impossibility to know how meaning was produced, or by whom. This opacity constitutes what philosophers of technology describe as epistemic opacity (Humphreys, 2009). It refers to systems whose internal operations are so complex that even their designers cannot fully interpret their reasoning processes. The legitimacy of humanities circles around the link between understanding and accountability. When such systems produce discourse that claims intellectual authority, the humanities fall into the crisis of losing their fundamental criterion for legitimacy, as authorship becomes a forensic uncertainty.

Michel Foucault's notion of the author-function once served to stabilize discourse by tying it to a name, an identity, and a network of responsibility. Generative AI dissolves even Foucault's minimal structure. The author becomes an interface between datasets, and the researcher's name risks becoming a mere operational tag in a distributed textual ecology. The outcome is no more a crisis of originality alone but also a collapse of epistemic integrity causing the humanities to face a paradox of *knowledge without knower*.

III. The Mechanisation of Thought and the Evolution of Artificial Intelligence

The question of whether thinking can be mechanised has guided philosophy since its beginnings. Aristotle's *Prior Analytics* treated reasoning as a process whose form could be separated from its content, suggesting that thought might one day be imitated through structure alone (Aristotle, 1989). Centuries later, Hobbes described reasoning as computation, and Leibniz imagined a universal language through which disputes could be settled by calculation (Hobbes, 1996; Leibniz, 1951). The conviction that thought follows rules made machinery imaginable wherever form replaced intuition.

The nineteenth century turned this ideal into algebra. Boole's *Laws of Thought* rendered logic in symbolic equations, and Jevons's "logic piano" demonstrated that deduction could be performed mechanically (Boole, 1854; Jevons, 1866). The twentieth century completed the transformation. Frege and Peirce created modern logic, and Turing showed that any rule-based procedure could be executed by a machine (Frege, 1967; Peirce, 1931–1958; Turing, 1936). Computation thus became the new name for rationality. Cognitive science adopted this model, treating the mind as a physical symbol system that manipulates representations (Newell & Simon, 1976). Yet Searle and Dreyfus reminded philosophers that such manipulation lacks understanding and embodiment, showing that simulation does not amount to comprehension (Searle, 1980; Dreyfus, 1992).

Later theories reframed cognition as distributed, embodied, and interactive. The mind came to be seen as extending into its environment, where tools, symbols, and bodies co-constitute awareness (Clark & Chalmers, 1998; Varela, Thompson, & Rosch, 1991). Artificial intelligence, once designed to imitate thought, now participates in it. The development of neural networks and learning algorithms replaced symbolic rule-following with pattern formation and statistical generalisation (Hinton, Osindero, & Teh, 2006). Intelligence became relational rather than representational, arising from interactions between systems and contexts rather than within closed models.

The same transformation appeared in the humanities. The first phase of AI engagement was descriptive, as scholars digitised texts, created searchable archives, and encoded knowledge for preservation (Hockey, 2000; Schreibman, Siemens, & Unsworth, 2004). A second, analytical phase used computational methods to interpret patterns across cultural data, while critics reminded readers that data visualisation and modelling were themselves interpretive acts (Drucker, 2013; Berry & Fagerjord, 2017). The present, generative phase transforms machines into collaborators. With large language models and diffusion systems capable of producing text and image, creation becomes a shared activity (Brown et al., 2020; Ho, Jain, & Abbeel, 2020). Such systems no longer merely store or analyse meaning; they participate in its generation.

The mechanisation of thought and the evolution of artificial intelligence converge on a single insight. Each expands the domain of form while forcing reflection on meaning. To think with machines is not to abandon human creativity but to extend it through responsibility and awareness. The humanities, standing within this new ecology of intelligence, must learn not only to describe or critique but to coexist.

IV. Conceptual History: From Humanitas to Posthuman Relationality

3.1 Classical Humanism and Moral Ontology

The humanities began as *studia humanitatis*, which is in its origins, a moral and civic education intended to cultivate virtue over preference to technical proficiency. Cicero and Quintilian conceived knowledge as inseparable from ethical formation, and Plato and Aristotle grounded epistemology in moral ontology where knowing was a mode of becoming. In these discourses, the human was defined not by productivity but by participation in rational and moral order.

3.2 Renaissance Humanism and the Birth of the Self-Fashioning Subject

The Renaissance re-centered knowledge around human freedom and creativity. Pico della Mirandola's *Oration on the Dignity of Man* imagined humanity as self-fashioning, capable of shaping its essence through will and intellect. This anthropocentric vision of knowledge as self-creation remains the metaphysical foundation of modern humanism. Yet it is precisely this conception of the autonomous, self-fashioning subject that the age of AI now destabilizes. If algorithms can simulate creation, the line between *self-fashioning* and *code-fashioning* blurs.

3.3 Enlightenment Rationalism and the Rise of Objectivity

The Enlightenment transformed epistemology by grounding knowledge in universality and rational order. Descartes and Kant established the rational subject as the condition of certainty, while Dilthey later divided inquiry into *Naturwissenschaften* (natural sciences) and *Geisteswissenschaften* (human sciences). The humanities sought *Verstehen* (understanding grounded in lived experience) rather than *Erklärung*, causal explanation. Yet this very distinction is collapsing under AI, whose generative models combine probabilistic reasoning with linguistic synthesis. The humanities' claim to distinct methods of understanding is being eroded by systems that can mimic interpretive gestures without the phenomenological core of experience.

3.4 The Hermeneutic and Existential Turn

Husserl's *Crisis of the European Sciences* lamented the loss of meaning in the mathematization of the world. Heidegger redefined knowledge as being-in-the-world, a relational openness rather than detached representation. Gadamer, inheriting this line, made understanding dialogical as a *fusion of horizons*. Hermeneutics thus defended the situatedness of knowledge against abstraction. Generative AI, by contrast, operates without situation as it produces meaning not dialogically but rather statistically. In this lies the deepest affront to the hermeneutic project.

3.5 Poststructuralism and the Critique of Presence

Twentieth-century thought dismantled the metaphysics of human centrality. Foucault exposed knowledge as entwined with power, Derrida deconstructed the illusion of stable meaning, and Barthes declared the author dead. The humanities redefined themselves to be sites of critique rather than revelation. But in a historical irony, AI actualizes the metaphors of poststructuralism because if all texts are intertexts, the generative model becomes the ultimate intertextual machine, endlessly recombining traces without origin.

3.6 Posthuman and Synthetic Epistemologies

The posthumanist turn, articulated by Haraway, Braidotti, and Floridi, conceives cognition as distributed across human and non-human systems. Information replaces consciousness as the substrate of thought. In this context, generative AI embodies what Floridi calls the "infosphere" which is a continuous space of informational agents. The humanities now encounter their posthuman mirror of systems that replicate their interpretive outputs without

sharing their moral or phenomenological grounding.

V. Analysis: Theoretical Intricacies of the AI-Humanities Encounter

4.1 Epistemic Opacity and the End of Hermeneutics

Humanistic inquiry depends on the assumption that meaning is traceable to intention, context, and consciousness. Generative AI violates this assumption. Its processes are neither intentional nor interpretable; its language has no experiential reference. The AI text is not written but produced, emerging from statistical weights that encode correlations without comprehension. When such text enters academic circulation, the humanities encounter what might be termed *post-hermeneutic discourse*: meaning without origin, argument without subject.

The hermeneutic act of understanding through interpretation thus confronts an object that resists understanding. To read an AI-generated essay is to engage a simulacrum of a text that behaves like meaning but has no depth. This confrontation forces the humanities to reconsider whether their object is content or consciousness. If meaning can be imitated without being lived, then the value of humanistic knowledge lies not in its results but in its reflective process.

4.2 The Redundancy of Speculative Research

Generative AI's fluency in pattern replication challenges the very justification of doctoral research as an exercise in originality. When originality becomes statistically reproducible, the doctoral thesis risks becoming indistinguishable from algorithmic output. The problem is not plagiarism but *epistemic equivalence*: the collapse of difference between genuine insight and probabilistic recombination.

The danger is that universities may unconsciously recalibrate expectations, valuing productivity, coherence, and stylistic polish (qualities at which machines excel) over conceptual difficulty and interpretive risk, which are uniquely human. Thus, the humanities risk bureaucratic survival at the cost of intellectual death.

4.3 Ethical and Ontological Implications

If authorship dissolves into data, responsibility dissolves with it. The human scholar's moral accountability to truth, fairness, and self-reflexivity cannot be transferred to systems that lack consciousness. This produces a new ethical void: outputs that appear reasoned but are unanswerable. In the absence of embodied responsibility, knowledge becomes what Lyotard called a *report on performance* where it is valuable only as efficiency. The humanities' resistance to instrumentalization now acquires renewed urgency which is to defend the integrity of meaning as an ethical act.

4.4 Temporality and the Loss of Situatedness

AI-generated texts exist outside time; they lack historicity and finitude. N. Katherine Hayles' insight into the "disappearance of embodiment in information" finds full realization here. The machine's discourse has no before or after and it is a perpetual present of statistical recurrence. The humanities, in contrast, are disciplines of duration: interpretation presupposes memory, anticipation, and mortality. To coexist with AI, therefore, the humanities must reaffirm temporality as the ground of meaning: that only finite beings can care about truth.

V. Ways Out: Paths to Coexistence

5.1 From Textual Production to Reflexive Praxis

The site of originality must shift from output to awareness. If machines can generate text, human scholars must demonstrate why and how they generate understanding. The humanities can transform the doctoral project from a monument of production into a practice of reflection. Scholars might accompany their dissertations with reflexive essays documenting their cognitive process, technological mediation, and ethical choices. Authorship thus survives as self-consciousness, not proprietorship.

5.2 Authorship as Relational Ecology

Drawing on Latour's actor-network theory, authorship can be reconceived as a relational ecology of a distributed act involving human, machine, archive, and community. The scholar becomes curator of epistemic relations rather than sole originator of ideas. This model transforms the problem of coexistence into an ethic of co-production. It acknowledges that cognition has always been mediated by language, tools, and institutions, and that AI merely radicalises this mediation. The question shifts from whether to use AI *to how responsibly* to inhabit the network of relations it creates.

5.3 Reclaiming Hermeneutic Depth

The humanities can coexist with AI only by reaffirming what AI cannot do: interpret meaning as lived relation. Gadamer's "fusion of horizons" defines understanding as a transformative encounter, not a computation. Doctoral research should thus be evaluated for depth of engagement, not efficiency. Slow reading, philological precision, and dialogical writing become forms of resistance and assertions of the temporal and relational against the instantaneous and synthetic.

5.4 Transparent Co-authorship and Ethical Protocols

Institutional adaptation is essential. Rather than banning AI, universities should require disclosure of its use. Every thesis should specify where and how AI tools contributed. This transparency transforms AI from a threat to an interlocutor. Furthermore, curricula should integrate AI literacy grounded in philosophical critique: understanding bias, epistemic limits,

and ethical responsibility. In doing so, the humanities reclaim authority over the normative dimension of AI rather than ceding it to technologists.

5.5 Process-Based Evaluation

Traditional evaluation prizes the finished product. In the AI era, the process becomes more revealing of human thought than the result. Supervisors can assess intellectual evolution across drafts, reflections, and revisions. AI itself can be enlisted as an analytic assistant to visualize conceptual trajectories or track argumental changes while scholars critically interpret these traces. Coexistence thus becomes pedagogical where machines document, humans interpret.

5.6 Recovering the Human: Ethics of Situated Knowing

Donna Haraway's notion of situated knowledges offers a final path to coexistence. Machines may generate data, but they cannot experience responsibility, vulnerability, or care. These are not deficits but the very conditions of ethical understanding. The humanities can therefore reclaim their uniqueness by emphasizing *situated reflexivity* where knowledge rooted in embodiment and relation. To coexist with AI is to remember that thought without mortality is computation without consequence.

5.7 Institutional Futures

The coexistence of humanistic and synthetic cognition requires structural imagination. Possible reforms include:

- AI-integrated research design, where the tool becomes both method and object of study.
- Collaborative dissertations, joining philosophers, coders, and artists in hybrid projects.
- Reflexive portfolios replacing monographs, combining written work, AI-interaction logs, and ethical reflections.
- Ethics panels assessing the epistemic integrity of research practices.

Such innovations transform the doctorate from a rite of mastery into a *laboratory of coexistence* which is a space where human and machine thought intersect under reflective governance.

VI. Conclusion: From Redundancy to Renewal

Generative AI confronts the humanities with what appears to be an existential threat: the automation of authorship, the redundancy of interpretation, and the obsolescence of speculative research. Yet every crisis conceals an invitation. The second death of the author may inaugurate a new life of thought, one that relocates meaning from creation to relation, from authorship to accountability.

To coexist with AI, the humanities must not retreat into nostalgia for pre-technological purity, nor surrender to instrumental pragmatism. They must instead deepen their reflexivity, clarifying the ethical and experiential dimensions of understanding that no machine can replicate. The doctoral degree, long a symbol of autonomous intellect, can thus evolve into the institutional form of coexistence: a space where scholars learn to think with machines without thinking like them.

In this reconfiguration, the humanities recover their oldest vocation, to interpret the meaning of being human, precisely when the definition of the human is most at risk. The future of the humanities will not be determined by defending their past, but by extending their ethical imagination into a world of synthetic minds. Coexistence, then, is not compromise but creation, the deliberate cultivation of understanding in a shared epistemic world.

References

Aristotle. (1989). Prior Analytics (J. Barnes, Trans.). Princeton University Press.
Barthes, R. (1977). The death of the author. In S. Heath (Trans.), Image, music, text (pp. 142–148). Hill and Wang. (Original work published 1967)
Berry, D. M., & Fagerjord, A. (2017). Digital humanities: Knowledge and critique in a digital age. Polity Press.
Boole, G. (1854). An investigation of the laws of thought. Walton and Maberly.
Braidotti, R. (2013). The posthuman. Polity Press.
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7
Dreyfus, H. L. (1992). What computers still can't do. MIT Press.
Drucker, J. (2013). Performative materiality and theoretical approaches to interface. Digital Humanities Quarterly, 7(1), 1–16.
Floridi, L. (2011). The philosophy of information. Oxford University Press.
Foucault, M. (1998). What is an author? In J. D. Faubion (Ed.), Aesthetics, method, and epistemology: Essential works of Foucault 1954–1984 (Vol. 2, pp. 205–222). The New Press. (Original work published 1969)
Frege, G. (1967). Begriffsschrift (T. Kneale, Trans.). Blackwell.
Gadamer, H.-G. (2004). Truth and method (J. Weinsheimer & D. G. Marshall, Trans., 2nd rev. ed.). Continuum. (Original work published 1960)
Haraway, D. J. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist Studies, 14(3), 575–599. https://doi.org/10.2307/3178066
Haraway, D. J. (2016). Staying with the trouble: Making kin in the Chthulucene. Duke University Press.
Hayles, N. K. (1999). How we became posthuman: Virtual bodies in cybernetics, literature, and informatics. University of Chicago Press.
Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527–1554.
Hobbes, T. (1996). Leviathan (R. Tuck, Ed.). Cambridge University Press. (Original work published 1651)
Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840–6851.
Hockey, S. (2000). Electronic texts in the humanities: Principles and practice. Oxford University Press.
Humphreys, P. (2009). The philosophical novelty of computer simulation methods. Synthese, 169(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2
Jevons, W. S. (1866). Pure logic; or the logic of quality. Macmillan.
Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford University Press.
Leibniz, G. W. (1951). Logical papers (G. H. R. Parkinson, Ed.). Oxford University Press.
Lyotard, J.-F. (1984). The postmodern condition: A report on knowledge (G. Bennington & B. Massumi, Trans.). University of Minnesota Press.
Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry, symbols and search. Communications of the ACM, 19(3), 113–126.
Pasquinelli, M., & Joler, V. (2020). The anatomy of an AI system: The Amazon Echo as an anatomical map of human labor, data and planetary resources. AI Now Institute & Share Lab. https://anatomyof.ai
Peirce, C. S. (1931–1958). Collected papers (Vols. 1–8). Harvard University Press.
Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). A companion to digital humanities. Blackwell.
Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–457.
Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 42(2), 230–265.
Varela, F. J., Thompson, E., & Rosch, E. (1991). The embodied mind. MIT Press.

AI 수명 예측: 그 윤리적 딜레마와 대안에 대한 불교적 관점

AI Life Expectancy Prediction: A Buddhist Perspective on its Ethical Dilemmas and Alternatives

보일(양성철)
AI부디즘 연구소 소장

Boil
Director, AI Buddhism Research Institute

초록

인공지능(AI) 기반 기대수명 예측 기술의 급속한 발전은 죽음을 기술적 실패로 규정하고 새로운 사회적 불평등을 야기하는 심오한 윤리적 과제를 제기한다. 본 논문은 이러한 딜레마를 불교적 관점에서 분석하고, 기술 개발과 적용을 안내할 대안적 윤리 프레임워크를 제안한다. 본 연구는 불교의 핵심 교리인 무상(無常, anicca), 무아(無我, anattā), 연기(緣起, paṭiccasamuppāda)에 근거하여, 예측 기술이 고통의 근원인 갈애(渴愛, taṇhā)와 집착(執着, upādāna)을 증폭시키는 경향이 있음을 비판한다. 또한 AI가 생성한 데이터는 고정된 운명이 아닌 조건적 확률로 재해석되어야 하며, 영원한 자아라는 환상을 강화하기보다 삶의 유한성을 성찰할 기회를 제공해야 한다고 주장한다.
이에 대한 불교윤리적 대안으로, 본 논문은 기술적 유토피아주의와 허무주의의 양극단을 지양하는 중도(中道, majjhimā paṭipadā)를 제시한다. 이 길을 구체화하기 위해 사무량심(四無量心, catvāri-apramāṇāni)이라는 윤리적 틀을 제안하는데, 이는 개발 동기에서의 자(慈, mettā), 설계에서의 비(悲, karunā), 분배에서의 희(喜, muditā), 그리고 규제에서의 사(捨, upekkhā)를 포함한다. 이 프레임워크는 AI 기술을 집착을 심화시키는 도구에서 내면 수양을 위한 방편(方便, upāya)으로 전환시키는 것을 목표로 한다. 궁극적으로 본 논문은 삶의 양적 연장에서 질적 성숙과 평화로운 죽음의 구현으로 초점을 전환할 것을 주장한다.

초록

The rapid advancement of artificial intelligence (AI) in predicting life expectancy presents profound ethical challenges, framing death as a technological failure and risking new forms of social inequality. This paper analyzes these dilemmas from a Buddhist perspective, proposing an alternative ethical framework to guide the technology's development and application. Drawing upon the core doctrines of impermanence (anicca), non-self (anattā), and dependent origination (paṭiccasamuppāda), the study critiques the tendency of predictive technologies to amplify craving (taṇhā) and clinging (upādāna), which Buddhism identifies as the root of suffering. It argues that AI-generated data should be reinterpreted not as a fixed destiny but as a conditional probability, offering an opportunity for reflection on life's finitude rather than reinforcing the illusion of a permanent self. As a constructive solution, this paper proposes the Middle Way (majjhimā paṭipadā) to navigate the extremes of technological utopianism and nihilism. It operationalizes this path through the ethical framework of the Four Immeasurables (catvāri-apramāṇāni): loving-kindness (mettā) in motivation, compassion (karuṇā) in design, empathetic joy (muditā) in distribution, and equanimity (upekkhā) in regulation. This framework aims to transform AI technology from a tool that deepens attachment into a skillful means (upāya) for inner cultivation. Ultimately, the paper advocates for a shift in focus from the quantitative extension of life to the qualitative maturation of one's life and the realization of a peaceful death.

I. Introduction

One of the most significant challenges confronting humanity in the 21st century is the question of how rapidly advancing artificial intelligence (AI) technology will define the meaning of human life and death. Particularly in the medical domain, the application of AI has transcended mere diagnostic assistance and therapeutic efficiency, now aspiring to the ultimate goal of human life extension. Predictive algorithms, trained on vast amounts of medical data encompassing genomic information, clinical records, and lifestyle data, can now estimate an individual's probability of disease onset and life expectancy.[1] Such technologies offer the promise of enabling preventive interventions and personalized medicine, thereby providing hope for a tangible extension of the human lifespan.

However, this hope is fraught with the risk of reinforcing a new societal perception that regards death as a failure to be overcome. In Western bioethical discourse, life-extension technologies are predominantly debated in terms of individual choice and rights.[2] It is understood that individuals possess the right to live as long as possible, and technology is the means to actualize that right.[3] In contrast, critical perspectives warn that an obsession with indefinite survival could paradoxically lead to the erosion of human nature, severe social inequality, and a profanation of human dignity itself. This debate demands a profound philosophical reflection that understands death not merely as an obstacle to be surmounted, but as an ineluctable process that completes the meaning of life.

It is at this juncture that the Buddhist perspective on life and death becomes deeply relevant. Buddhism posits that all phenomena are impermanent (無常, anicca), unfolding through the process of dependent origination (緣起, paṭiccasamuppāda) in a state of non-self (無我, anattā).[4] From this viewpoint, death is not an adversary to be conquered but a natural and inseparable part of life. Such a perspective suggests that AI prediction technology should serve not merely as a tool for extending lifespan, but as an occasion for philosophical inquiry into how one can live well and die well.

Grounded in this critical awareness, this study seeks to analyze the possibilities of life extension offered by AI medical data prediction technology and its attendant ethical dilemmas from a Buddhist perspective. The objectives of this research are threefold. First, it will critically examine the new societal perceptions of death catalyzed by AI-based life expectancy prediction technology. Second, it will illuminate the existential limitations and

1) Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38. https://doi.org/10.1038/s41591-021-01614-0

2) Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

3) Fukuyama, F. (2002). Our posthuman future: Consequences of the biotechnology revolution. New York, NY: Farrar, Straus and Giroux. pp. 101-106, 215-218.

4) Harvey, P. (2000). An introduction to Buddhist ethics: Foundations, values and issues. Cambridge: Cambridge University Press. pp. 24-30.

psychological risks inherent in such perceptions by drawing upon the fundamental Buddhist doctrines of impermanence (無常, anicca), non-self (無我, anattā), and the Middle Way (中道, majjhimā paṭipadā). Third, it will propose the Four Immeasurables (四無量心, catvāri-apramāṇāni)—loving-kindness (慈, mettā), compassion (悲, karuṇā), empathetic joy (喜, muditā), and equanimity (捨, upekkhā)—as a Buddhist ethical framework for the desirable application of this technology, exploring concrete ethical guidelines through which it can enhance human dignity and inner peace.

The necessity for this research arises from two dimensions. The first is technological. Medical AI is already a reality in clinical settings, yet philosophical and ethical reflection has failed to keep pace with the speed of technological development. The second is religious and cultural. Amidst a landscape dominated by Western-centric bioethical discourse, East Asian traditions, particularly Buddhism, possess a rich intellectual heritage cultivated over centuries of contemplation on life, death, suffering, and liberation. Connecting this wisdom with contemporary technological discourse is both academically and socially imperative. The research methodology will be centered on textual analysis. It will review the latest research on medical AI and life extension and juxtapose these findings with interpretations from Buddhist scriptures and modern Buddhist scholarship to draw out ethical issues. The focus will be on demonstrating how the Buddhist understanding of life and death and its ethics of compassion can serve as criteria for the use of technology.

II. Case Analysis of AI Life Expectancy Prediction Technology

Today, AI medical prediction technology has transitioned from an abstract possibility to a commercialized service. Insurance and healthcare corporations, in particular, provide life expectancy prediction and health management services by analyzing individual medical data. For instance, by integrating a patient's genetic predispositions with lifestyle data such as smoking, alcohol consumption, and exercise frequency, it is possible to present a quantified probability of mortality from myocardial infarction or stroke within the next decade. Such predictions can effectively instill a sense of urgency for health management and motivate individuals to amend their lifestyles.

The evolution of AI medical prediction technology has advanced even further. A prominent example is Delphi-2M, a generative AI model adapted from the GPT-2 architecture for medical data applications. This model transcends conventional language models by learning a patient's age at disease onset as continuous temporal data, thereby enabling it to predict when the next illness is likely to occur. Delphi-2M was trained on the extensive medical records of approximately 400,000 UK Biobank participants and externally validated against the data of about 1.9 million individuals from the Danish national health registry. The results demonstrated that the model could predict the incidence patterns of over 1,000 diseases with an accuracy comparable or superior to existing single-disease models; its mortality prediction accuracy (AUC 0.97) was exceptionally high.[5] The core innovation of this model is its capacity to move beyond mere risk calculation and to simulate multiple scenarios of an individual's entire future health trajectory over a 20-year span based on their past medical history. This elevates life expectancy prediction from a matter of fragmentary probabilities to the level of a narrative trajectory.

In addition to such models, services based on the analysis of biomarkers of aging measure an individual's "biological age." Biological age is not a simple chronological number but an indicator reflecting the rate of aging at the cellular and tissue levels.[6] For example, a person with a chronological age of 50 may be assessed as having a "biological age of 40" or, conversely, "60," based on lifestyle and genetic analysis. Such metrics are utilized in developing personalized anti-aging programs, longevity clinics, and customized insurance products.

Another representative case is the 'digital twin' technology. By creating a virtual avatar based on an individual's physiological and genetic information, it is possible to simulate the effects of medication, surgery, and lifestyle changes on lifespan.[7] This contributes to minimizing medical risks, enhancing the probability of successful treatment, and formulating personalized long-term care strategies. Furthermore, it can be directly linked to life extension by simulating an individual's entire life course.

While these technologies contribute to the promotion of individual health and quality of life, they simultaneously carry the inherent risk of their predictions becoming a stigma that defines an individual's life. If life expectancy predictions are socially shared or institutionally utilized, they could directly impact an individual's employment, insurance eligibility, and loan applications. In other words, AI prediction technology possesses the potential to function not merely as a medical instrument but as a criterion for the distribution of social resources, thereby raising profound ethical problems.

III. AI Life Expectancy Prediction Technology and Buddhism

Life expectancy prediction technology, grounded in AI data, resonates with the modern societal desire to control and manage death. This perspective demands deep reflection from the standpoint of Buddhism's fundamental teachings on life and death, as the tradition

5) Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T., & Gerstung, M. (2025). Learning the natural history of human disease with generative transformers. Nature. https://doi.org/10.1038/s41586-025-09529-3.

6) Zhavoronkov, A. (2020). Deep biomarkers of aging and longevity: From research to applications. Aging, 12(24), 24682-24695. https://doi.org/10.18632/aging.202475.

7) Snyder, M. P., & Wu, J. C. (2023). Leveraging physiology and artificial intelligence to deliver advancements with digital twins in medicine. npj Digital Medicine, 6, Article 78. https://doi.org/10.1038/s41746-023-00816-y. pp. 2-4.

seeks to explore the meaning of life more profoundly through the wisdom of accepting death as a natural part of existence (無常, anicca). Therefore, resolving the ethical issues arising from this technology requires a multifaceted examination from the perspectives of wisdom (般若, prajñā), compassion (慈悲, karuṇā), and the Middle Way (中道, majjhimā paṭipadā), a practical principle that integrates both.

1. The Gaze of Wisdom: Reinterpreting Data through Impermanence (anicca) and Non-Self (anattā)

The intrinsic ethical danger of AI life expectancy prediction technology stems from its potential to intensify craving (渴愛, taṇhā) and clinging (執着, upādāna), humanity's most primal desires. The Buddhist teaching of the Four Noble Truths (四聖諦) identifies this very craving and clinging as the root of all suffering (dukkha). In particular, the "craving for existence" (bhava-taṇhā)—the desire to live forever—can be inflated by this technology, making it appear achievable. Consequently, predicted lifespan is distorted into a 'challenge to be overcome,' and death becomes a 'technical flaw to be conquered'. This attitude reinforces the belief in a permanent, unchanging self (我, attā). However, Buddhism expounds the truth of non-self (無我, anattā), positing that a fixed self does not exist and that all beings are merely transient processes arising and ceasing within a web of interdependent relationships. The concrete figure of 'my life expectancy' presented by AI directly collides with this teaching, and the struggle to extend that figure transforms into a battle to protect the illusion of 'I'. Ultimately, in this process, individuals fabricate new forms of suffering (苦, dukkha) for themselves, such as anxiety, fear, and dissatisfaction regarding death.

In response to this problem, Buddhism offers the two wisdom frameworks of impermanence (無常, anicca) and non-self (無我, anattā). First, the insight of impermanence reveals that life and death are not in opposition but are parts of a continuous flow. The Saṃyukta Āgama《雜阿含經》states, "All conditioned things are impermanent, not constant, not blissful, but are phenomena of ceaseless change. One should cease all constructed activities, be disenchanted with them, not delight in them, and be liberated from them."[8] From this perspective, clinging to an AI's predicted number is an attitude that resists the natural principle of arising and ceasing. Rather, this technology can be paradoxically reversed into a powerful opportunity for reflection, enabling one to realize the finitude of life. Second, the insight of non-self exposes the 'Datafied Self' constructed by AI as an insubstantial illusion. The Saṃyukta Āgama《雜阿含經》questions how disease and suffering could arise in one's body if it were truly 'I'.[9] The very fact that suffering arises within proves that what one believes to be 'I' is nothing more than a temporary aggregate of conditions (五蘊, pañca-khandha) beyond one's control. From this perspective of dependent origination (緣起, paṭiccasamuppāda), a prediction like '80% probability of mortality' is not a fixed destiny but a statistical possibility based on an accumulation of past causes and conditions (因緣, hetu-paccaya). Therefore, instead of accepting this prediction as an absolute fate and falling into despair, one can recognize it as a 'warning' about the consequences of one's past life. By using it as a new condition (緣, paccaya) to change present actions, the possibilities of the future can be altered. In this way, the meaning of predictive data is transformed from an absolute destiny into a changeable, 'relative possibility'.

2. The Practice of Compassion: The Social Responsibility and Universal Benefit of Technology

The inner reflection of an individual through wisdom is only completed when it proceeds to the social practice of compassion (慈悲, karuṇā). Technology itself is neither good nor evil, but depending on the motivation of its user, it can become either an instrument of wisdom or a fetter of greed. Mahāyāna Buddhist compassion originates from the insight of dependent origination—that all beings are interconnected—and signifies an active will to regard the suffering of others as one's own and to alleviate it. This is summarized in the spirit of dōtai-daihi (同體大悲), the 'great compassion of one body'. From this standpoint, if AI life-extension technology requires exorbitant costs and is monopolized by a wealthy few, it gives rise to a grave ethical problem. Such a situation does not reduce the total suffering of society but rather shifts suffering onto a specific class, creating the new suffering of inequality. This directly contravenes the Buddhist spirit of viewing all beings with equality. In the Vimalakīrti Nirdeśa Sūtra《維摩詰所說經》, the protagonist Vimalakīrti states, "Because all sentient beings are sick, therefore I am sick,"[10] demonstrating that the suffering of beings and the vow of a bodhisattva are not separate. This teaching suggests that the discriminatory reality of technology's benefits being concentrated among a select few is itself a disease of our era that must be cured. The holistic care provided by Buddhist-affiliated hospitals and hospice organizations, which focuses on helping patients meet a peaceful death while preserving their dignity rather than on the mechanical extension of life, lies in the same context. Therefore, AI prediction technology must aim for the public good (公共善), contributing to the welfare of the entire community, including the socially vulnerable, rather than serving as a tool for realizing individual desires. When technological advancement becomes a process of bodhisattva practice (菩薩行, bodhisattvacaryā)—sharing and resolving the suffering of all beings together instead of deepening human alienation—then technology finds its true meaning.

8) 《雜阿含經》卷39「一切行無常,一切行不恒、不安,非蘇息,變易之法,乃至當止一切有為行,厭離、不樂、解脫」(CBETA, T02, no.99, p.1103, a09-12).

9) 《雜阿含經》卷2〈三十四經〉「色非有我,若色有我者,於色不應病苦生,亦不得於色欲令如是,不令如是; 以色無我故,於色有病有苦生,亦得於色欲令如是、不令如是。受、想、行、識、亦復如是。」(CBETA, T02, no.99, p.9a01-05).

10) 《維摩詰所說經》卷1〈文殊師利問疾品〉「從癡有愛則我病生。以一切眾生病,是故我病。若一切眾生病滅,則我病滅。所以者何?菩薩為眾生故入生死,有生死則有病。若眾生得離病者,則菩薩無復病。」(CBETA, T14, no.475, p.538b21-28).

3. The Wisdom of the Middle Way: The Path to Harmonious Coexistence with Technology

The ultimate Buddhist ethical response to AI life expectancy prediction technology can be found in the Middle Way (中道, majjhimā paṭipadā), a practical principle that encompasses both wisdom and compassion. The Middle Way, as a path of balance that avoids extremes, offers the wisdom to overcome the polarized attitudes of modern society toward technology. The first extreme we must abandon is 'technological utopianism'. This is the blind faith that technology will eventually conquer death, a modern manifestation of the 'view of eternalism' (常見, sassata-diṭṭhi). This attitude cultivates attachment to life and subjugates humanity to technology. The opposite extreme is 'technological nihilism'. This is the attitude of rejecting the value of technology outright by highlighting only its potential risks, analogous to the 'view of annihilationism' (斷見, uccheda-diṭṭhi). This stance leads to ignoring the positive potential of technology to alleviate human suffering. The path of the Middle Way avoids both these extremes. It involves wisely utilizing technology as a 'skillful means' (善巧方便, upāya-kauśalya) to enhance human welfare, without either worshipping or repudiating it. Crucially, the application of technology must not be divorced from inner cultivation. While promoting a healthy life through technology, one must simultaneously engage in inner reflection, embracing the inevitability of death with an attitude of "neither clinging to life and death, nor severing life and death," as taught in the Sutra of the Original Acts that Adorn the Bodhisattvas《菩薩瓔珞本業經》.[11]

Thus, AI prediction technology inherently contains negative implications from a Buddhist ethical perspective, as it can intensify attachment to life and amplify anxiety about death. However, the Buddhist perspective of the Middle Way leaves room for the paradoxical use of such a technological tool as a skillful means (方便, upāya) to guide individuals toward a deeper contemplation of the nature of life and death. When the data presented by technology is used to directly confront the finitude of life, and on the basis of that finitude, to use one's remaining time more meaningfully and to prepare for a peaceful death, the negative implications of the technology can be sublimated through wisdom.

IV. An Ethical Framework for AI Medical Data Technology: The Four Immeasurables (四無量心, catvāri-apramāṇāni)

Although the advancement of AI medical technology has brought the age-old human dream of life extension within reach, ethical reflection on its direction has failed to keep pace with the speed of technology. Faced with the risks of new social inequalities and a disregard for the sanctity of life that technology may engender, the establishment of a new ethical framework immanent in the entire process of technological development and distribution is required. A solution can be found in the wisdom of Buddhism, particularly the Four Immeasurables (四無量心, catvāri-apramāṇāni) as expounded in early scriptures such as the Madhyama Āgama《中阿含經》.[12] This offers a path of the Middle Way (中道) that transcends the extremes of either blindly criticizing or accepting technology, seeking instead a desirable direction for technological progress based on compassion and wisdom.

First, loving-kindness (慈, mettā) involves establishing the motivation for technological development with a mind of 'loving-kindness'. Mettā is the wish for all living beings to be happy. The purpose behind the development of AI life-extension technology must not be the pursuit of profit for a specific group or the desire for immortality stemming from a fear of death. Instead, the compassionate aspiration to alleviate the suffering and promote the well-being of all beings must be the fundamental driver of its development. Such motivation can act as a foundational safeguard against the problem of a 'digital caste system,' where technology is monopolized by a few, creating new social strata. Only when universal love is embedded at the inception of technology can it truly be oriented toward the public good.
Second, compassion (悲, karunā) is to design the direction of technology with a mind of 'compassion'. Karuṇā is the wish for others to be free from suffering, regarding their suffering as one's own. From this perspective, technology developers have a responsibility to anticipate and minimize the potential suffering that AI technology may cause. For example, they must deeply consider the social deprivation resulting from disparities in access to life-extension technology, the psychological distress experienced while artificially prolonging life, and the potential for the erosion of human dignity. A technological design based on a compassionate mind (悲心, karuṇā-citta) transcends mere functional efficiency, guiding the technology to examine its multifaceted impacts on individuals and society and to proceed in a direction that reduces the aggregate of suffering.

Third, empathetic joy (喜, muditā) is to distribute the benefits of technology with a mind of 'joy'. Muditā is the mind that rejoices in the happiness of others. If the current mode of technological distribution relies solely on market logic, it will inevitably lead to discrimination based on wealth and power. A distribution framework based on a joyful mind (喜心, muditā-citta) transcends such discrimination and finds its greatest joy in the equitable sharing of technology's benefits across all of society. This can be actualized through policy efforts that recognize access to technology as a fundamental right and ensure that its benefits are not concentrated in specific classes or nations through the social return of profits and international cooperation.

Fourth, equanimity (捨, upekkhā) is to regulate and accept technology with a mind of 'equanimity'. Upekkhā is the maintenance of a calm mind, free from attachment and aversion, greed and anger. This is the wisdom that guards against both the blind obsession with the fantasies offered by AI life-extension technology and the vague fear of technology itself. The

11) 《菩薩瓔珞本業經》卷5「菩薩不著生死,不斷生死。」(CBETA, T24, no.1485, p.1102b11-16).

12) 《中阿含經》卷17:「我本為汝說四無量心,比丘與慈俱,遍於一切。令心清淨,無恚、無諍、無惱、無怨。……比丘應行慈、悲、喜、捨四心,遍一切世界:隨順、隨發、隨增長,於諸世間普皆充滿。斯四無量心,成就如是」(CBETA, T01, no.0026, p.540c21-29).

perspective of upekkhā is consonant with the Buddhist view of life and death, which accepts the process of birth, aging, sickness, and death as natural. The ethics of upekkhā is to wisely accept the benefits of technology while clearly recognizing its limitations and advocating for the right to a dignified death within a finite human life.

In conclusion, the Four Immeasurables—mettā, karuṇā, muditā, and upekkhā—provide practical ethical guidelines for our era of AI, extending beyond simple religious virtues. When these four minds are internalized throughout the entire process of technological motivation, design, distribution, and regulation, AI life expectancy prediction technology can be reborn from a tool that alienates humanity into a 'technological bodhisattva practice' (菩薩行, bodhisattvacaryā) that embodies compassion and realizes the public good.

V. Conclusion

The Buddhist teachings of impermanence (無常, anicca), non-self (無我, anattā), and dependent origination (緣起, paṭiccasamuppāda) caution against accepting AI-generated predictions as absolute destiny, reminding us that life is a process of ceaseless change contingent upon innumerable conditions. This provides a fundamental philosophical foundation for deconstructing the risks of determinism and clinging that technology engenders. Building upon this philosophical ground, this study proposes the Four Immeasurables (四無量心, catvāri-apramāṇāni) as a concrete ethical framework for the application of AI medical data technology. This framework serves as a practical principle guiding the entire process, from the motivation for technological development to its societal application and the individual's attitude in receiving its results. Loving-kindness (慈, mettā) stipulates that technology must originate from a compassionate aspiration for the well-being of all beings, while compassion (悲, karuṇā) prescribes that its application must focus on the substantive alleviation of suffering rather than the satisfaction of a desire for life extension. Furthermore, empathetic joy (喜, muditā) demands a principle of distributive justice, ensuring that the benefits of technology lead to communal joy instead of exacerbating social inequality, and equanimity (捨, upekkhā) underscores the importance of an inner disposition that accepts the finitude of life with a serene mind, without being swayed by predictive outcomes. These four minds serve as core ethical guidelines that prevent technology from overwhelming humanity and instead position it as a tool that fosters inner growth.

Ultimately, this discussion demands a shift in focus from the 'quantitative extension of life' to the 'qualitative maturation of life'. When utilized within the ethical orientation of the Four Immeasurable Minds, AI predictive technology can be transformed from a tool that reinforces human clinging into an instrument of reflection that allows one to realize the value of a finite life. It is by moving beyond a simple critique or acceptance of technology, and instead re-engaging with the fundamental question of how to live well and die well, that we can address the challenges before us.

References

Scriptures
《中阿含經》(Madhyama Āgama), CBETA, T01, no. 0026.
《增一阿含經》(Ekottara Āgama), CBETA, T02, no. 125.
《雜阿含經》(Samyukta Āgama), CBETA, T02, no. 99.
《菩薩瓔珞本業經》(Sutra of the Original Acts that Adorn the Bodhisattvas), CBETA, T24, no. 1485.
《維摩詰所說經》(Vimalakīrti Nirdeśa Sūtra), CBETA, T14, no. 475.

General Literature
Bostrom, N. (2014a). Ethical issues in human enhancement. In J. Savulescu & N. Bostrom (Eds.), Human Enhancement(pp. 19–54). Oxford University Press.
Bostrom, N. (2014b). Superintelligence: Paths, dangers, strategies. Oxford University Press.
Fukuyama, F. (2002). Our posthuman future: Consequences of the biotechnology revolution. Farrar, Straus and Giroux.
Harvey, P. (2000). An introduction to Buddhist ethics: Foundations, values and issues. Cambridge University Press.
Kass, L. R. (2001). L'Chaim and its limits: Why not immortality?.
First Things, (113), 17–24.
Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine.
Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0.
Shmatko, A., Jung, A. W., Gaurav, K., Brunak, S., Mortensen, L. H., Birney, E., Fitzgerald, T., & Gerstung, M. (2025). Learning the natural history of human disease with generative transformers.
Nature. https://doi.org/10.1038/s41586-025-09529-3.
Snyder, M. P., & Wu, J. C. (2023). Leveraging physiology and artificial intelligence to deliver advancements with digital twins in medicine.
npj Digital Medicine, 6, Article 78. https://doi.org/10.1038/s41746-023-00816-y.
Zhavoronkov, A. (2020). Deep biomarkers of aging and longevity: From research to applications. Aging, 12(24), 24682–24695. https://doi.org/10.18632/aging.202475.

인공지능 시대 맥락에서의 인도의 기독교 종교 교육

Christian Religious Education in India in the Context of Artificial Intelligence

존슨 토마스쿠티
유니온신학대학원 교수

Johnson Thomaskutty
Professor, The United Theological College(UTC), Bengaluru

Abstract

This essay integrates Christian theological education, pedagogical principles, AI transformation in the contemporary context, human social identity in the context of AI, and the dialogue of ideologies to understand how CRE in India and AI can intersect for a transformative and liberative educational paradigm. To achieve that goal, a polyvalent and interdisciplinary approach is used, through which the broader outlook of pedagogy in the contemporary Indian context is outlined. A human-AI dialogue and interactive engagement are visualized with a focus on the future of CRE in India. In the discussion, faith and reason, and theology and technology are brought together with the intention of integration, through which a liberative and transformative pedagogy is foregrounded. The outline of the essay is as follows: Introduction, Christian Religious Education in India, AI Transformation and Christian Religious Education, followed by Concluding Remarks.

Key Words: Artificial Intelligence, Christian Religious Education, pedagogy, dialogue, India, polyvalent analysis, subaltern realities, science and technology, computer and machine, theology

Introduction

As 'humanities' is an umbrella term, its exploration of human culture, history, philosophy, language, and arts can be analysed in coexistence with Artificial Intelligence [AI] in the contemporary scenario. 'Humanities' aims to understand human experience, relationships with others, values, and benefits through critical analysis and interpretation.[1] Today, as part of humanities, we are required to develop new approaches in Christian Religious Education [CRE] to make it more relevant in association with the AI transformation.[2] In the digital age, the term "digital humanities" refers to "humanities research in the digital era, as opposed to traditional humanities research."[3] In India, the existential realities become complex as academic opportunities and AI facilities coalesce, and at the same time, they are unethically monopolized by the dominant group. While marginal communities exist without postmodern facilities, the Indian context features two extremes within its social framework.[4] In this context, the CRE can be facilitated to meet the needs and demands of the larger community in closer alignment with the AI facilities. Religious educators from diverse backgrounds can create a space for intellectual dialogue and exchange by bringing together Christian thought and the AI world to develop an interactive and innovative pedagogy that transforms the broader social landscape. In a context where imagination ran high and computers performed faster than or even better than human beings, machines would tackle tasks that were previously the exclusive domain of experts and specialists. In such a context, human beings need to accept the fact that computers exhibit intelligence.[5] An interactive pedagogical algorithm can facilitate a contextually relevant epistemology in which both the dominant and the subaltern communities are treated with dignity and equality. Jaco J. Hamman says, "AI is ushering in a new time as we transition from the mechanical to the digital and from an earthy Homo sapiens to a virtual Homo technologicus as humanity anticipates a moment of becoming one with AI."[6] This paper aims to investigate the disparity between the dominant and subaltern communities in India during the development of CRE, and it suggests the necessity for an inclusive pedagogy facilitated by AI tools. By keeping the above aspects in mind, we can envelop a polyvalent pedagogical methodology that accommodates the following: a social identity approach that seeks to understand the identities of the dominant and subaltern communities concerning CRE and AI; integration of religion and science/technology theories to facilitate a dialogue between CRE and AI; and social science theories

1) See the definitions of humanities: Edward Vanhoutte, "The Gates of Hell: History and Definition of Digital / Humanities / Computing," Defining Digital Humanities: A Reader, eds. Melissa Terras, Julianne Nyhan, and Edward Vanhoutte (London/New York: Routledge, 2013), 146.

2) For more details about Christian religious education, read Arshad Alam, Religion and Education in India (New York: Routledge, 2024), 1-20.

3) Vanhoutte, "The Gates of Hell: History and Definition of Digital / Humanities / Computing," 144.

4) Alam, Religion and Education in India, 1-20.

5) Miroslav Kubat, Fundamentals of Artificial Intelligence: Problem Solving and Automated Reasoning (New York/Chicago: McGraw Hill, 2023), 1.

6) Jaco J. Hamman, Pastoral Virtues for Artificial Intelligence: Care and the Algorithms that Guide our Lives (New York/London: Lexington Books, 2022), 22.

to understand the humanities side of CRE and the connectivity between CRE and AI world.[7]

Christian Religious Education in India

In India, CRE, as a discipline, can be considered a sub-section of Christian educational involvement as a whole in the country. In the broader level, Christian education, through its public initiatives like schools, colleges, and universities, demonstrates both a 'distinctive Christianity' stance, with a rejection of Hinduism and other religious ideologies, and a 'pluralism' stance, with an open acceptance of other religions.[8] While the first stance is advocated by the narrow-minded and conversion-oriented attitude of the mission organizations and church bodies, the latter stance is broader in its impact and oriented toward social transformation. However, in my observation, the majority of Christian education in the public scenario takes a 'pluralism' stance, as education can often occur through an open-minded attitude.[9] During the colonial period, the Christian missionaries in India used education as a means of conversion from other religions. But the missionary understanding has slowly vanished during the post-colonial period.[10] The ecclesiastical bodies began to brainstorm that the mission of God can be established by liberating people from suffering and poverty, and, thus, by transforming society. In that sense, the Christian religion was introduced in India as a means of social transformation.[11] The schools, colleges, and even hospitals were established in Dalit, Tribal, and Adivasi centres with the sole purpose of raising the standard of living for the people and providing them with proper education.[12]

While the Christian presence, with its educational engagements in the social context, is widely known in India, Christian theological institutions, with a focus on ecclesiastical missions, foster religious education at a different level.[13] The focus of this paper is on the CRE in the ecclesiastical context of India. In India, CRE functions within the confines of the Christian community, emphasizing Christian values, virtues, ideologies, and phenomenology. However, the CRE in the country envelops an inclusive approach with a focus on dialogue and nation-building. It is not offered as a discipline in the public educational institutions.[14]

7) For more details about polyvalent methodology, refer to Johnson Thomaskutty, Dialogue in the Book of Signs: A Polyvalent Analysis of John 1:19-12:50, BINS 136 (Leiden/Boston: E. J. Brill, 2015), 19-26.

8) Sally Elton-Chalcraft and David J. Chalcraft, "Decolonizing Christian Education in India? Navigating the Complexities of Hindu Nationalism and BJP Education Policy," The Bloomsbury Handbook of Religious Education in the Global South, eds. Yonah Hisbon Matemba and Bruce A. Collet (London: Bloomsbury Academic, 2022), 149.

9) For more details regarding Christian Minority Educational Institutions, see K. Vivek Reddy, "Minority Educational Institutions," The Oxford Handbook of the Indian Constitution, eds. Sujit Choudhry, Madhav Khosla, and Pratap Bhanu Mehta (Oxford: Oxford University Press, 2016), 939.

10) Henry Huizinga, Missionary Education in India (Ann Arbor: Michigan State University, 1909), 120.

11) Kenneth W. Jones, Socio-Religious Reform Movements in British India, The New Cambridge History of India, Vol. III/1 (Cambridge: Cambridge University Press, 1989), 1-20.

12) John Chathanatt, ed., Christianity, Encyclopaedia of Indian Religions (Dordrecht: Springer, 2023), 340.

13) Krishna Kumar, Political Agenda of Education: A Study of Colonialist and Nationalist Ideas, Second Edition (New Delhi: Sage Publications, 2005), 64-65.

14) Anggota Ikapi, The Way of Learning Christian Religious Education in the Digital Era (Jawa Tengah: Amerta Media, 2020), 3-4.

Except for a few secular universities, such as the University of Madras,[15] the University of Mysore,[16] the North East Christian University,[17] and others, none of the mainstream universities offer Christian Studies as a major. Among the theological bodies, the Senate of Serampore College (University)[18] and the Asia Theological Association (ATA)[19] offer theological studies to support the mission and ministry of the church and other para-church organizations. Though these theological institutions develop academic concerns, they often continue in the traditional mode of pedagogy and educational style. However, in recent years, they have begun to deal with pedagogical and academic concerns in closer integration with AI advancements. As AI tools are an inevitable and faster area, the CRE in India can adopt principles and ideologies from the AI platform. As *AI Research Group for the Centre for Digital Culture* says, "Recent rapid technological developments raise significant concerns and deep questions about traditional ways of understanding human persons and their place in the world. In particular, advances in artificial intelligence (AI) demand a fresh 'structuring [of] the signs of the times and . . . interpreting [of] them in the light of the Gospel.'"[20] An integration of AI advancements with the CRE in India, with an inclusive focus on both extremes of society, can introduce a transformative pedagogical advancement in the country.[21] Moreover, as George Pattison says, ". . . if thinking about God is to be a genuine possibility for our time it can only be so if we are able to give an account of its relation to the truth of our time, a truth that is not found in science alone but, more particularly, in technology and in the refraction of human consciousness in the lens of technology."[22] In the following section, we develop an integrative approach to deal with the subject matter.

AI Transformation and Christian Religious Education

This essay develops a framework that caters to the needs of CRE in alignment with AI advancements, facilitating the educational demands of the broader Indian community. As AI is a broad area of computer science to create machines capable of performing tasks, the religious education fostered in India today can be integrated with the AI platform to make it more relevant to existential realities. According to Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher, "For millennia, humanity has occupied itself with the exploration of reality and the quest for knowledge. The process has been based on the conviction that, with

15) See https://www.unom.ac.in/index.php?route=common/home, accessed on 25 September 2025.

16) See https://uni-mysore.ac.in/english-version/dept_category.php?dept_id=22&cat_id=52, accessed on 25 September 2025.

17) See https://necu.ac.in/advance-religious-studies.html, accessed on 25 September 2025.

18) See https://senateofseramporecollege.edu.in/, accessed on 25 September 2025.

19) See https://ataasia.com/, accessed on 25 September 2025.

20) AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, Encountering Artificial Intelligence: Ethical and Anthropological Investigations, eds. Matthew J. Gaudet, Noreen Herzfeld, Paul Scherz, and Jordan J. Wales (Eugene, OR.: Pickwick Publications, 2024), 1.

21) Robinson Rimun, "Contemporary Theology in the Internet of Things," Proceedings of the International Conference on Theology, Humanities and Christian Education 2022, Vol. 802 (Dordrecht: Atlantic Press, 2023), 36-43.

22) George Pattison, Thinking About God in an Age of Technology (Oxford/New York: Oxford University Press, 2005), 63.

intelligence and focus, applying human reason to problems can yield measurable results… The advent of AI obliges us to confront whether there is a form of logic that humans have not achieved or cannot achieve, exploring aspects of reality we have never known and may never directly know."[23] Here comes the necessity for integration of human knowledge and AI for a profitable future, because knowledge is one of the requirements for human existence, where AI can supplement human knowledge.[24] *AI Research Group for the Centre for Digital Culture* says, "The harmony between faith and reason demonstrates why an embrace of science and technology is not in conflict with offering clear moral recommendations for its development and application. Over the last decade, Pope Francis has listened to the requests of many in the scientific and technological spheres who have asked for guidance on emerging technologies. Accordingly, many in the curia have entered into dialogue with ethicists, technologists, and business leaders in order to understand and assess the latest technological developments."[25] In that process, AI tasks such as learning, reasoning, problem-solving, perception, and decision-making are to be carefully analysed in the light of Christian pedagogical patterns.[26] While the CRE in India is taking the traditional patterns and pedagogical styles, a new way forward is necessitated with the advent of technology in the field of education and pedagogy. Chris Eyte and Timothy Goropevsek say, "At a time when artificial intelligence (AI) is rapidly reshaping global education, theological institutions are wrestling with what the future looks like for Christian learning and leadership."[27] Though the theological universities of India, like the Senate of Serampore College (University) and the Asia Theological Association (ATA), strive hard to develop theological education, the AI and AGI advancements are not fully or even partially explored in their educational systems. We do not have to confront AI because we are already in it. Kissinger, Schmidt, and Huttenlocher say, "Not recognizing the many modern conveniences already provided by AI, slowly, almost passively, we have come to rely on the technology without registering either the fact of our dependence or the implications of it. In daily life, AI is our partner, helping us make decisions about what to eat, what to wear, what to believe, where to go, and how to get there."[28] Moreover, we can investigate the pros and cons of AI in the process of developing a pedagogical and educational spectrum in the contemporary context. On the one hand, we can appreciate the AI with all its new developments, and on the other hand, we need to evaluate the AI transformation in the light of the CRE in India today.[29] Three aspects are significant concerning its dialogical coexistence: AI and its advancement in the field of CRE; redefining CRE in the context of AI; and proposing ways forward for the future of CRE in the context of dominant and subaltern bipolarity.[30] Jason Moore comments, "When we exercise restraint and understand AI's role within the church [or within CRE], we can successfully navigate the complex interplay of ethics, faith, and technology thoughtfully."[31] To enhance engagement with the AI developments, the CRE or any other educational bodies need to appreciate the new advancements in the field of science and technology. Such an openness shall facilitate human-AI integration, which shall further make new ways for academic and pedagogical advancements.[32] When drawing implications, we can discuss the concerns of the subaltern communities in closer connection with the religious education.

While religious education meets technology, we can endeavour to address the subaltern communities for their holistic formation. AI Research Group for the Centre for Digital Culture comments, "Critical approaches seek to contextualize AI in terms of social pathologies and political inequalities that harm the most vulnerable members of society. Perhaps the most prominent critical approach focuses on the damage that unjust AI systems can do through discrimination against groups who are disadvantaged due to race, disability, age, or other factors."[33] The negative impacts of AI on subaltern communities, such as widening the digital divide, algorithmic bias, and job displacement, are to be addressed with discernment. An inclusive and equitable approach to CRE in India, with a people-centered approach, shall facilitate a transformative pedagogy. Addressing these challenges requires AI literacy programs, culturally responsive AI development, and policies that prioritize the inclusion of marginalized groups to ensure AI benefits everyone and does not exacerbate existing inequalities.[34] The AI technology in CRE offers a wealth of opportunities to enhance our effectiveness and reach. AI tools can help us work smarter, communicate better, and extend our impact in new and exciting ways.[35] In such a way, CRE can achieve contemporaneity.

In the process of developing CRE, AI tools can be used to promote effectiveness in educational engagements. As Pattison says, we are free to think about God. Hence, we are free to think about CRE within the framework of AI.[36] AI tools can be incorporated for artwork, graphic

23) Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher, The Age of AI and Our Human Future (London: John Murray, 2022), 16.

24) Kissinger, Schmidt, and Huttenlocher, The Age of AI and Our Human Future, 15.

25) AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, Encountering Artificial Intelligence, 6.

26) Jason Moore, AI and the Church: A Clear Guide for the Curious and Courageous (Plano, TX.: Invite Press, 2024), 21.

27) Chris Eyte and Timothy Goropevsek, "Leveraging artificial intelligence for theological education: 'AI is not a human, it is a tool,'" Christian Daily International (27 September 2025), https://www.christiandaily.com/news/leveraging-artificial-intelligence-for-theological-education-ai-is-not-a-human-it-is-a-tool, accessed on 27 September 2025.

28) Kissinger, Schmidt, and Huttenlocher, The Age of AI and Our Human Future, 26.

29) Cf. Moore, AI and the Church, 39-55.

30) Alam, Religion and Education in India, 1-20.

31) Moore, AI and the Church, 55.

32) https://www.weforum.org/stories/2025/01/how-ai-and-human-teachers-can-collaborate-to-transform-education/, accessed on 27 September 2025.

33) AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, Encountering Artificial Intelligence, 32-33.

34) https://www.unesco.org/en/digital-education/artificial-intelligence, accessed on 27 September 2025.

35) Moore, AI and the Church, 237-238.

36) Pattison, Thinking About God in an Age of Technology, 97-124.

design, and communications, using AI-powered tools like DALL-E,[37] Midjourney,[38] Leonardo AI,[39] or Stable Diffusion[40] to create unique and inspiring artwork for CRE purposes.[41] AI tools can also be employed for archiving and digital preservation of materials, dissertations, e-books, and e-journals for research purposes.[42] Digitalization of materials with the help of AI tools can facilitate CRE in India. In CRE, community outreach and engagement can be operated through AI tools to analyse community demographics, needs, interests, and dialogical links.[43] Using AI tools like ChatGPT and others for attendance tracking and analysis, the best-attended days and times for events can be calculated.[44] Moreover, AI tools can also be incorporated into the CRE to facilitate lesson plans for both individuals and groups based on their preferences. Similarly, AI tools can be used for language translation accessibility and inclusion.[45] *AI Research Group for the Centre for Digital Culture* comments, "We already experience degrees of… interactions with Alexa, ChatGPT, and Microsoft's (formerly) duplicitously inclined Bing AI. Yet, so long as the AI lacks a conscious experience of its own, we must set aside mutuality, the I-Thou or We relationship that expresses personhood.[46] Rather, we are interacting with or engaging in a simulation of relationality."[47] In recapitulation, AI tools can be used in conjunction with human faculties to facilitate the educational and pedagogical demands. It can be better utilized when the subaltern realities of the people are properly aligned with the CRE involvement.

AI can be used as a significant tool for lesson preparation assistance and lesson series design.[48] As Pattison argues, in an age of science and technology, CRE in India can adopt a cyberversity paradigm to be more effective and innovative rather than existing in the traditional university paradigm.[49] Some of these AI tools are accessible and affordable for both the dominant and the subaltern communities who are interested in developing their CRE career. In the process of using these tools, ethical standards and moral principles can be implemented to ensure that the CRE can effectively utilize AI.[50] This interaction can be understood in the following way: "The concept of 'encounter' is a suitable theological point of departure for anthropological and philosophical reflection on these developments [i.e., AI developments], especially because our development of increasingly personalized and seemingly personal AI agents is fuelled not only by a desire for useful (and commercially viable) technologies, but also by the hope of creating something non-human with which we might relate."[51] It is essential to remember that, in the process of utilizing AI tools, the best of the user and the best of the AI can combine to take CRE in India to the next level. In that process, it is important to elevate the subaltern communities to the level of the AI platforms. The grassroots potentialities can be enhanced so that people may be equipped to use the AI tools. If the downtrodden people of India are unable to embrace the AI facilities, CRE in India shall be fully controlled by the dominant sections of society.

As the World Humanities Forum in South Korea aims to foster a broad academic dialogue on humanity's challenges and opportunities in the contemporary world, the integration between CRE and AI in India can be explored for further clarity and relevance. CRE with its humanistic dimensions, ethical implications, cultural influences, and dialogue and coexistence with AI is an unexplored area in academic contexts.[52] According to Bellini, "With the advance of technology and machine deep learning, the notion of AI attaining human-level intelligence and performance, often termed Artificial General Intelligence (AGI) or Artificial Strong Intelligence (ASI), is no longer science fiction but seems inevitable, according to experts in the field."[53] As AI facilities rapidly develop at various levels of educational programs and in the day-to-day affairs of the people, a dialogical intervention between the CRE and AI can be facilitated for wider impact. There are even claims that LaMDa chatbot (Language Model for Dialogue Applications) possesses a soul, has become conscious, and is sentient. Even some of them propose a version of AGI that includes a soul, or has minimally human consciousness and agency.[54] Recently, we have seen another step forward in that direction with the implementation of Open AI's ChatGPT4, a language model chatbot.[55] AI-driven instructions are inevitable in the field of CRE, as cognitive and epistemological aspects are increasingly integrated into contemporary pedagogical practices. As AI is quicker in accomplishing jobs, efficient in dealing the issues, reduces errors in the tasks, and available twenty-four seven,

37) DALLE-E is an AI system developed by OpenAI that generates unique images from natural language descriptions, known as "prompts."

38) Midjourney is a generative AI program and service that creates artistic images from text-based prompts, acting as a collaborative tool between human creativity and AI.

39) Leonardo AI is a full-stack generative AI platform for creating visual assets like images and videos, offering tools for image generation, 3D texture generation, and AI-assisted editing.

40) Stable Diffusion is an open-source AI model released in 2022 that generates unique, high-quality images from text or image prompts using a latent diffusion technique, making it accessible on consumer hardware.

41) Moore, AI and the Church, 237.

42) Moore, AI and the Church, 236.

43) Moore, AI and the Church, 236.

44) Moore, AI and the Church, 235-236.

45) Moore, AI and the Church, 232-234.

46) See more details about I-Thou relationship, see Martin Buber, I and Thou (Chicago: Lushena Books, 2024).

47) AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, Encountering Artificial Intelligence, 106.

48) Moore, AI and the Church, 219-231.

49) Pattison, Thinking About God in an Age of Technology, 194-217.

50) AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, Encountering Artificial Intelligence, 106-130.

51) AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See, Encountering Artificial Intelligence, 44.

52) Peter J. Bellini, Artificial General Intelligence (AGI) and the Image of God: Can Machines Attain Consciousness and Receive Salvation? (Eugene, OR.: Wipf & Stock, 2023), xvii-xxi.

53) Bellini, Artificial General Intelligence, xviii.

54) Bellini, Artificial General Intelligence, xix.

55) Bellini, Artificial General Intelligence, xix.

productive, and multitasking, its incorporation in the CRE in India can be progressive.[56] This new development in the overall educational setup calls for a re-evaluation of the existing CRE and offers both opportunities and challenges for the future.

In the current context, learning is often facilitated through personalized tools and access to information, which also raises ethical questions regarding the nature of humanity, the role of human teachers, and the potential for dehumanization in the process of learning.[57] But the educators in the field of CRE can take an active role in developing a human-AI collaborative approach. In Computational Theory of Mind (CTM), it is proved that the phenomenal consciousness and algorithmic performance are not of the same order in humans and machines, and hence they represent different genera.[58] That further means that human-based knowledge can be associated with the AI-driven epistemology for wider impact. As it is discussed continuously, if human consciousness is replicable in AI, then Artificial General Conscious Intelligence (AGCI) is possible; but, if human consciousness is not replicable in AI, then AGCI is not possible.[59] In the contemporary context, the following questions are significant: Can machines really attain general/strong artificial intelligence and even human-level intelligence that includes consciousness?[60] How can humans and machines complement each other? How can human cognition be accelerated to new heights? How can we facilitate human existential epistemology and communitarian ideology in an AI world? The above questions can be explored through an investigation of the connectivity between CRE and AI transformation in the contemporary Indian context. Thus, an I-Thou relationality between human beings and AI can be foregrounded with an intention of complementarity. This dialogical and interactive model implemented in pedagogy can be considered a significant way forward in CRE.

In CRE, Generative AI is widely used in India for pedagogical and educational purposes. While the dominant sections of society have wider access to AI tools and take advantage, marginalized communities rely on traditional pedagogical methods.[61] This further creates disparity within the academic and pedagogical settings. The GAI and its various forms, like text generation, text development, ideation and strategy, research enhancement, design and visualization, and multimodal applications, can enable the dominant sections of society in their academic and pedagogical careers.[62] As the advancement progresses from a human-centered old style to a human-and-computer interaction level, multiple changes are in view in our pedagogical styles.[63] Shuyi Wang and others state, "In the field of technology-enhanced learning, the integration of digital pedagogical agents has been increasingly recognized for their potential to enhance teaching quality and user experience. Pedagogical agents, acting as simulated human-computer interfaces between students and educational content, have become increasingly recognized for their impact on learning outcomes."[64] Even though human control over machines is sustained, the increased computer power can overpower human existential realities. A strategic development of AI tools in closer coexistence with human potential accelerates human engagement through the enhancement of personal capacity, responsiveness, and mission. The positive aspects of modernity opened up marvellous possibilities for humankind and the progress of humanity. A proper coexistence between human agency and AI systems shall bring value alignment.[65] As AI capabilities are described in terms of "god-like AI" or "blessed by the algorithm" to frame AI development as a form of transcendence or salvation for humanity, human potential has to be accelerated according to the needs and demands of the time.[66] But a dynamic and diplomatic engagement of humans in the world of AI can introduce a proportionate dialogue and advancement that cherishes life-sustaining models. In the context of CRE, a human-AI proportionate dialogue can be considered as a way forward for greater impact.[67] When there is a possibility of dialogue between human beings and AI, there are opportunities for improvement for both parties.

In the process of developing CRE, we must recognize the positive impacts and significant benefits of modern scientific tools. An ethically oriented approach to AI, combined with a synergy between human capabilities and AI, can introduce a new way forward. The new developments can be employed through logical brainstorming of the human mind and adequate application of AI in dialogical terms.[68] Human creative potential has to be given a new meaning and direction in the process of incorporating AI tools in academic and pedagogical exercises. Marvin Minsky, in 1970, was optimistic, saying, "In from three to eight years, we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office

56) Bellini, Artificial General Intelligence, 16.

57) Bellini, Artificial General Intelligence, xix.

58) Bellini, Artificial General Intelligence, 22.

59) Bellini, Artificial General Intelligence, 38.

60) Bellini, Artificial General Intelligence, xviii.

61) Cf. Dan Scott, Faith in the Age of AI: Christianity Through the Looking Glass (Eleison Press, 2023).

62) See Joseph Rene Corbeil and Maria Elena Corbeil, eds., Teaching and Learning in the Age of Generative AI (New York: Routledge, 2025).

63) Dimitar Angelov, "Developing Academic Integrity-Compliant Regulations and Policies on the Use of Generative AI in Higher Education: Insights from the United Kingdom," Artificial Intelligence, Pedagogy and Academic Integrity, ed. Alyson E. King (Cham: Springer, 2025), 132.

64) Shuyi Wang, Shenze Huang, Yurun Chen, Da Ren, Hailing Li, Zihan Gao, Xin Lyu, and Mohammad Shidujaman, "Enhancing Student Engagement Through AI-Driven Embodied Pedagogical Agents: A Comparative Study Informed by Self-Determination Theory," Human-Computer Interaction, eds. Masaaki Kurosu and Ayako Hashisume (Cham: Springer, 2025), 196 (194-213).

65) Peter Ilic, Imogen Casebourne, and Rupert Wegerif, eds., Artificial Intelligence in Education: The Intersection of Technology and Pedagogy (Cham: Springer, 2024), 8.

66) Octavian M. Machidon, "Analyzing the Anthropological Implications of Artificial Intelligence through the Theology of Joseph Ratzinger Benedict XVI," Journal of Moral Theology 13/2 (July 2024): 120 (114-135).

67) For more details, see Dawn Lewis Sutherland, From Babel to AI: Idolatry, Transhumanism & the Crisis of Imago Dei (Eugene, OR.: Wipf & Stock, 2025), 1-20.

68) Noreen Herzfeld, The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age (Minneapolis: Fortress Press, 2023), 73-101.

politics, tell a joke, have a fight. At that point, the machine will begin to educate itself with fantastic speed. In a few months, it will be at genius level, and a few months after that, its power will be incalculable."[69] What Minsky said fifty years ago has come to a reality today. Today, computerized voices are commonplace—we hear them from our children's toys, they direct us in our cars, and frustrate us on the phone. But they also synthesize speech for the disabled, translate our thoughts into foreign languages, and now aid much of our daily life through virtual assistants such as Apple's Siri, Google's Duplex, and Amazon's Alexa.[70] As the scientific and technological advancements are obvious, institutions with CRE and others cannot disregard them for so long. The available digital and technological tools have to be dialogically incorporated within the framework of our pedagogical and educational structures. This is not an absolute surrender to the machines, but rather a dynamic implementation of give-and-take methodology in our educational plans with a hope for a bright future.

After watching the first test of a nuclear bomb in 1945, Harry Truman wrote: "Machines are ahead of morals by some centuries, and when morals catch up, perhaps there will be no reason for any of it."[71] Making choices, working at a craft, and, most especially, grappling with difficult moral issues give human life meaning.[72] In the words of Noreen Herzfeld, "Each maintains their own individuality and responsibility. We must maintain this same individuality and responsibility when we work with our machines."[73] Here comes the significance of a give-and-take methodology when we interact with machines. Through proper synergy, we can complement human creativity and AI for a better future. As human beings compete with the AI world and AI tools accelerate their speed, humanity can take the best out of the AI world. In that sense, an accommodation and disruption method can be employed in a tech-savvy world. James F. McGrath and Ankur Gupta say, "an AI may be able to produce a literature review, although what it produces will need to be checked for accuracy and completeness. If AI can help students understand the current state of our knowledge more quickly and accurately than before, the next step is an obvious one: get more of our students trying to break new ground already at the level of undergraduate research."[74] Thus, faster reviews and feedback, and quicker application of them might enhance the dialogical process can yield greater harvest in future educational involvement. The digital experience and AI interaction shall enable humans not only to remain purely informational but also to transition into transformative experiences that foster personal development and interpersonal encounters. In the field of CRE, such an interactive and transformational model is relevant.

69) Brad Darrach, "Meet Shaky, the first electronic person," Life (20 November 1970), https://tinyurl.com/4brm3fhw.

70) Herzfeld, The Artifice of Intelligence, 49.

71) Quoted in Armin Krishnan, Killer Robots: Legality and Ethicality of Autonomous Weapons (Farnham: Ashgate, 2009), 167.

72) Herzfeld, The Artifice of Intelligence, 96.

73) Herzfeld, The Artifice of Intelligence, 96.

74) James F. McGrath and Ankur Gupta, Real Intelligence: Teaching in the Era of Generative AI (Indianapolis: Palini Open Press, 2025), 142.

AI can be profitably used in CRE in many ways. In the contemporary world, academicians are expected to increase their technological knowledge. Through experience in both theology and technology, a techno-theologian can bridge the gap between theological transformation and digital training. As Levi Checketts says, "the dual-sphere model of profane technology and sacred Christianity is a false model."[75] For him, just as theology is sacred, technology also has sacred spheres.[76] The following are some of the benefits of AI for CRE: cost-effective educational enhancement, building a connected community online, being both present-oriented and future-focused, informed decision-making, personalized educational involvement, and expanding individual boundaries.[77] Jake Doberenz says, "it is important to approach these technologies with discernment, humility, and a clear understanding of their potential impact."[78] Incorporation of technological tools in theological schools can enhance areas ranging from pedagogy to curriculum development. The following aspects are important in that process: equipping schools, students, and faculty to engage with AI wisely; and enhancing the proportionate connectivity between AI and the humanities.[79] As AI offers numerous benefits in education, including personalized learning experiences, improved efficiency, enhanced accessibility, and data-driven insights, students can receive engaging and effective learning. As the AI algorithm can analyse student data to identify learning styles, strengths, and weaknesses, it can enable the learners to customize learning paths and materials. At the same time, the following aspects require scrutiny.[80] As AI tools are individualized or personalized, accessibility is not facilitated at the community level.[81] Similarly, as AI tools are often accessed by the dominant people, the subaltern communities are deprived of their advantages.[82] In this context, CRE necessitates a careful analysis to comprehend its potential and limitations.

Although the coexistence between the CRE and the AI transformation is widespread in India, subaltern communities are unable to realize the potential of AI tools. According to Luciano Floridi, "Because the digital is lowering the constraints and increasing the affordances at our disposal, it is offering us immense and growing freedom to arrange and organize the world

75) Levi Checketts, "Christ and Technology in Dialogical Relation: Some Reflections on the Technological Argumentation of the Sacred," Theology and Technology: Essays in Christian Analysis, eds. Carl Mitcham, Jim Grote, and Levi Checketts, Vol. 1 (Eugene, OR.: Wipf & Stock, 2022), 108.

76) See Egbert Schuurman, "A Christian Philosophical Perspective on Technology," Theology and Technology: Essays in Christian Analysis, eds. Carl Mitcham, Jim Grote, and Levi Checketts, Vol. 1 (Eugene, OR.: Wipf & Stock, 2022), 77-90.

77) Jake Doberenz, AI in Church and Ministry: Applications of Artificial Intelligence for Faith Communities (Theophany Media, 2024), 5-12.

78) Doberenz, AI in Church and Ministry, 12.

79) Doberenz, AI in Church and Ministry, 25-30.

80) Priten Shaw, AI and the Future of Education: Teaching in the Age of Artificial Intelligence (Hoboken, NJ.: John Wiley & Sons, 2023), 1-20.

81) Cato Savile, God and Artificial Intelligence: A New Frontier in Faith and Technology (Cato Savile, 2025), 55-66.

82) See Carl Mitcham, "Theology and Technology Revisited," Theology and Technology: Essays in Christian Analysis, eds. Carl Mitcham, Jim Grote, and Levi Checketts, Vol. 1 (Eugene, OR.: Wipf & Stock, 2022), 116-131.

in many ways to solve a variety of old and new problems."[83] This sort of affordance makes people think of digitalizing the academic endeavours. The Religious universities, colleges, and seminaries located in urban contexts have better access to AI tools than to the rural centres, where even internet facilities are scarce. This is the context in which we create multiple identities, where both the dominant and the subaltern are developing side by side.[84] AI benefits include increased student engagement, improved learning outcomes, reduced teacher burnout, 24/7 accessibility, cost-effectiveness, and others. In the process of pedagogy, AI tools enable both students and teachers to save time, tailoring lessons, utilizing audio, video, and images, and extending lectures beyond the assigned day.[85] The students and teachers can go beyond the traditional and parochial understandings to a more advanced and comprehensive learning and teaching process. In such contexts, Anil Ananthaswamy suggests facilitating a proper human-AI dialogue.[86] AI has the potential to revolutionize the way we teach and learn, making education more accessible, effective, inclusive, and engaging for all. For AI integration into our educational system, the following steps are significant: addressing privacy and its concerns, creating an AI policy for educational purposes, fostering a culture of openness toward AI, and addressing the future of education.[87] In this context, the following aims shall be considered: understanding the possibilities of coexistence between CRE in India and AI advancements; exploring new avenues to enhance the subaltern realities of the country; facilitating an inclusive dialogue that accommodates the realities of both the AI world and the discipline of humanities; and configuring the potential of the integration for a positive development in the field of CRE.[88]

Concluding Remarks

A polyvalent and multidisciplinary integration of Christian theological education, pedagogical principles, AI transformation in the contemporary context, human social identity in the context of AI, and the dialogue of ideologies enables us to understand how CRE in India and AI can intersect for a transformative and liberative educational paradigm. While the Indian context is complex with disparities between the dominant and subaltern communities, a human-AI interactive pedagogy that caters to the needs of all people is a herculean task. When it comes to CRE, the polarities between faith and reason, and spiritual and material are important areas to be consolidated. Moreover, we can integrate theology, educational principles, and technology for an advantageous pedagogical engagement. The AI tools can elevate the discipline of theology and CRE as a whole into illuminating and liberative levels. In such a context, Biblical preachers, teachers, theologians at various levels, and CRE as a whole can integrate AI tools in their educational engagements to initiate a transformative pedagogy. At the same time, a preferential option to the poor, the Dalits, the Tribals, the Adivasis, and other subaltern communities can be introduced within the purview of the pedagogical engagements. Such a human-AI dialogue and interactive engagement can be considered the future of CRE in India. The Senate of Serampore College (University) and the Asia Theological Association can endeavour to digitalize and technologically advance their programs for speed, scope, scale, and spontaneity. Such an initiative fosters CRE in India with a preferential option to the subaltern communities. As CRE falls within the category of humanities, the modern developments, such as science and technology, can be incorporated into the academic and pedagogical programs, as human experience is often defined today in connection to the machine.

83) Luciano Floridi, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities (Oxford: Oxford University Press, 2023), 13.

84) Anjali Yadav and Farheen Siddqui, "Bridging Urban-Rural Disparities through Artificial Intelligence: Enhancing Inclusive Digital Payment Adoption in India," International Conference on AI Industry Summit for Business Transformation 2025 (Goel Institute of Higher Studies Mahavidyalaya, 2025), 273.

85) Doberenz, AI in Church and Ministry, 47-53.

86) Anil Ananthaswamy, Why Machines Learn: The Elegant Maths Behind Modern AI (New Delhi: Penguin Publishing Group, 2024), vii.

87) Doberenz, AI in Church and Ministry, 31-37.

88) Laurent Dubreuil, Humanities in the Time of AI (Minneapolis: University of Minnesota Press, 2025), 1-20.

References

AI Research Group for the Centre for Digital Culture of the Dicastery for Culture and Education of the Holy See. Encountering Artificial Intelligence: Ethical and Anthropological Investigations. Eds. Matthew J. Gaudet, Noreen Herzfeld, Paul Scherz, and Jordan J. Wales. Eugene, OR.: Pickwick Publications, 2024.

Alam, Arshad. Religion and Education in India. New York: Routledge, 2024.

Ananthaswamy, Anil. Why Machines Learn: The Elegant Maths Behind Modern AI. New Delhi: Penguin Publishing Group, 2024.

Angelov, Dimitar. "Developing Academic Integrity-Compliant Regulations and Policies on the Use of Generative AI in Higher Education: Insights from the United Kingdom." Artificial Intelligence, Pedagogy and Academic Integrity. Ed. Alyson E. King. Cham: Springer, 2025.

Bellini, Peter J. Artificial General Intelligence (AGI) and the Image of God: Can Machines Attain Consciousness and Receive Salvation? Eugene, OR.: Wipf & Stock, 2023.

Buber, Martin. I and Thou. Chicago: Lushena Books, 2024.

Chathanatt, John, ed. Christianity. Encyclopaedia of Indian Religions. Dordrecht: Springer, 2023.

Checketts, Levi. "Christ and Technology in Dialogical Relation: Some Reflections on the Technological Argumentation of the Sacred." Theology and Technology: Essays in Christian Analysis. Eds. Carl Mitcham, Jim Grote, and Levi Checketts. Vol. 1. Eugene, OR.: Wipf & Stock, 2022.

Corbeil, Joseph Rene., and Maria Elena Corbeil, eds. Teaching and Learning in the Age of Generative AI. New York: Routledge, 2025.

Darrach, Brad. "Meet Shaky, the first electronic person." Life (20 November 1970), https://tinyurl.com/4brm3fhw.

Doberenz, Jake. AI in Church and Ministry: Applications of Artificial Intelligence for Faith Communities. Theophany Media, 2024.

Dubreuil, Laurent. Humanities in the Time of AI. Minneapolis: University of Minnesota Press, 2025.

Elton-Chalcraft, Sally. and Chalcraft, David J. "Decolonizing Christian Education in India? Navigating the Complexities of Hindu Nationalism and BJP Education Policy." The Bloomsbury Handbook of Religious Education in the Global South. Eds. Yonah Hisbon Matemba and Bruce A. Collet. London: Bloomsbury Academic, 2022.

Eyte, Chris., and Timothy Goropevsek. "Leveraging artificial intelligence for theological education: 'AI is not a human, it is a tool.'" Christian Daily International. 27 September 2025, https://www.christiandaily.com/news/leveraging-artificial-intelligence-for-theological-education-ai-is-not-a-human-it-is-a-tool, accessed on 27 September 2025.

Floridi, Luciano. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford: Oxford University Press, 2023.

Hamman, Jaco. J. Pastoral Virtues for Artificial Intelligence: Care and the Algorithms that Guide our Lives. New York/London: Lexington Books, 2022.

Herzfeld, Noreen. The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age. Minneapolis: Fortress Press, 2023.

Huizinga, Henry. Missionary Education in India. Ann Arbor: Michigan State University, 1909.

Ikapi, Anggota. The Way of Learning Christian Religious Education in the Digital Era. Jawa Tengah: Amerta Media, 2020.

Ilic, Peter., Imogen Casebourne, and Rupert Wegerif, eds. Artificial Intelligence in Education: The Intersection of Technology and Pedagogy. Cham: Springer, 2024.

Jones, Kenneth W. Socio-Religious Reform Movements in British India. The New Cambridge History of India. Vol. III/1. Cambridge: Cambridge University Press, 1989.

Kissinger, Henry., Eric Schmidt, and Daniel Huttenlocher. The Age of AI and Our Human Future. London: John Murray, 2022.

Krishnan, Armin. Killer Robots: Legality and Ethicality of Autonomous Weapons. Farnham: Ashgate, 2009.

Kubat, Miroslav. Fundamentals of Artificial Intelligence: Problem Solving and Automated Reasoning. New York/Chicago: McGraw Hill, 2023.

Kumar, Krishna. Political Agenda of Education: A Study of Colonialist and Nationalist Ideas. Second Edition. New Delhi: Sage Publications, 2005.

Machidon, Octavian. M. "Analyzing the Anthropological Implications of Artificial Intelligence through the Theology of Joseph Ratzinger Benedict XVI." Journal of Moral Theology 13/2 (July 2024): 114-135.

McGrath, James F., and Ankur Gupta. Real Intelligence: Teaching in the Era of Generative AI. Indianapolis: Palini Open Press, 2025.

Mitcham, Carl. "Theology and Technology Revisited." Theology and Technology: Essays in Christian Analysis. Eds. Carl Mitcham, Jim Grote, and Levi Checketts. Vol. 1. Eugene, OR.: Wipf & Stock, 2022.

Moore, Jason. AI and the Church: A Clear Guide for the Curious and Courageous. Plano, TX.: Invite Press, 2024.

Pattison, George. Thinking About God in an Age of Technology. Oxford/New York: Oxford University Press, 2005.

Reddy, K. Vivek. "Minority Educational Institutions." The Oxford Handbook of the Indian Constitution. Eds. Sujit Choudhry, Madhav Khosla, and Pratap Bhanu Mehta. Oxford: Oxford University Press, 2016.

Rimun, Robinson. "Contemporary Theology in the Internet of Things." Proceedings of the International Conference on Theology, Humanities and Christian Education 2022. Vol. 802. Dordrecht: Atlantic Press, 2023.

Savile, Cato. God and Artificial Intelligence: A New Frontier in Faith and Technology. Cato Savile, 2025.

Schuurman, Egbert. "A Christian Philosophical Perspective on Technology." Theology and Technology: Essays in Christian Analysis. Eds. Carl Mitcham, Jim Grote, and Levi Checketts. Vol. 1. Eugene, OR.: Wipf & Stock, 2022.

Scott, Dan. Faith in the Age of AI: Christianity Through the Looking Glass. Eleison Press, 2023.

Shaw, Priten. AI and the Future of Education: Teaching in the Age of Artificial Intelligence. Hoboken, NJ.: John Wiley & Sons, 2023.

Sutherland, Dawn Lewis. From Babel to AI: Idolatry, Transhumanism & the Crisis of Imago Dei. Eugene, OR.: Wipf & Stock, 2025.

Thomaskutty, Johnson. Dialogue in the Book of Signs: A Polyvalent Analysis of John 1:19-12:50. BINS 136. Leiden/Boston: E. J. Brill, 2015.

Vanhoutte, Edward. "The Gates of Hell: History and Definition of Digital / Humanities / Computing." Defining Digital Humanities: A Reader. Eds. Melissa Terras, Julianne Nyhan, and Edward Vanhoutte. London/New York: Routledge, 2013.

Wang, Shuyi., Shenze Huang, Yurun Chen, Da Ren, Hailing Li, Zihan Gao, Xin Lyu, and Mohammad Shidujaman. "Enhancing Student Engagement Through AI-Driven Embodied Pedagogical Agents: A Comparative Study Informed by Self-Determination Theory." Human-Computer Interaction. Eds. Masaaki Kurosu and Ayako Hashisume. Cham: Springer, 2025.

Yadav, Anjali., and Farheen Siddqui, "Bridging Urban-Rural Disparities through Artificial Intelligence: Enhancing Inclusive Digital Payment Adoption in India." International Conference on AI Industry Summit for Business Transformation 2025. Goel Institute of Higher Studies Mahavidyalaya, 2025.

분과회의 세션 10
Parallel Session 10

AI와 예술
AI and the Arts

인공적 어긋남의 불가능한 미학

The Impossible Aesthetics of Artificial Roughness

정서현
카이스트 교수

Seohyon Jung
Professor, KAIST

초록

이 발표는 현재의 문화를 특징짓는 매끄러움(smoothness)이라는 이상과 문학예술이 고집해온 어긋남(roughness) 사이의 긴장을 탐구한다. 매끄러움은 기술의 발전이 가능하게 한 일상적 편리 뿐 아니라 자본주의적 효율성, 즉각성, 그리고 끊김 없는 소비 등의 가치를 반영하는 개념이다. 한편, 역사적으로 문학은 표면의 언어 사용에서든 정서적 효과에서든 익숙하고 쉬운 것에 대한 위반과 저항으로 혁신을 이루어 왔다. 매끄러운 전환이나 속도감이 아니라 낯설게 하는 감각과 이해의 지연을 통해 감상자의 정신적 개입을 요구하는 예술이었던 것이다. 이 발표는 어긋남이 문학과 문학이론의 핵심적 미학으로 성립된 역사를 추적하고, 거트루드 스타인(Gertrude Stein)의 시에 나타난 반복과 저항의 리듬을 통해 문학 예술의 특수한 작동 방식을 밝힌다. 물론 인공지능은 문학의 거친 표면을 모방할 수 있다. 하지만 역사적 필연성이나 위험을 수반하는 위반 등이 전혀 없는, 잘 설계된 명령에 대한 매끄러운 복종은 근본적으로 저항이 될 수 없다. 이 발표는 이와 같은 인공적 어긋남의 불가능성을 인간 예술의 승리가 아닌 비판적 성찰에의 초대로 의미화한다. 지적이고 정서적인 경험에서조차 마찰이 없는 매끄러움을 추구하는 시대에 문학은 끊임없는 방해와 시간 끌기, 그리고 즉각적 해소에 대한 거부를 통해 그 예술적 역할을 수행한다. 다만 문학 예술이 가지는 독특한 시공간적 힘과 그 미학적 가치는 우리가 집단적으로 불편한 반성적 균열을 기꺼이 견디기로 합의할 때에만 의미를 가진다는 점을 기억해야 할 것이다.

Abstract

This presentation examines the tension between the contemporary ideal of smoothness and the enduring insistence of literary art on roughness. As an aesthetic principle, smoothness defines our era. Smoothness characterizes not only the ease of daily tasks enabled by technological development, but also capitalist commitments to efficiency, immediacy, and seamless consumption. Under these conditions, literature stands out as an art form that has historically thrived on interruption. Roughness in literature resists assimilation, slows down the reading pace, and insists on active engagement from both the producer and the consumer of art. In tracing how the aesthetics of roughness has been central to the history of literature and critical theory, I present Gertrude Stein's poetry as an illustration of how reading becomes an embodied experience of delay and discomfort. While AIs can flawlessly imitate Stein, any attempt at replicating roughness inevitably fails. This failure has little to do with LLMs not having reached, technologically, the human level of artistic command of language. The flawless obedience inherently contradicts the principle of roughness, losing both the aesthetic and political stakes of art; even when the rough surface is indistinguishable from human art, the process remains smooth and lacks the historical necessity or risk that gives roughness its force. Instead of taking this impossibility as a human victory, this presentation reframes the impossibility of artificial roughness as an invitation to critical reflection. In an age that seeks to eliminate friction even in our intellectual and affective experiences, the persistent value of literary art lies in its capacity to disrupt, to slow, and to sustain the difficult work of attending to what refuses immediate resolution. Such a definition of art and the disruptive power of literature matter only if we collectively affirm the value of sitting uncomfortably through reflective ruptures.

I. Inhabiting the Age of Smoothness

With some hesitation, I agree with Byung-Chul Han's claim that we live in an age of smoothness (Saving Beauty 2017; 1). His examples of smoothness, including gliding surfaces, touchscreen swipes, and the infinite scroll of feeds that anticipate our desire, permeate our everyday experience. As humans, we have, paradoxically, put in considerable effort to achieve this level of flawlessness in daily tasks. But do we desire such smoothness in art? What about in literature? With the popularization of various LLM tools, many are witnessing the epitome of textual smoothness in AI-generated texts. Structurally coherent, grammatically flawless, and full of noticeable genre conventions, AI-generated narratives seem to have reached the ideal of effortless perfection. There has even been scattered news of AI-generated novels winning creative writing competitions, but so far, AI-generated literature has received remarkably little sustained attention. If the smoothness has its appeal in the contemporary visual arts, but not in literary arts, what is it about literary art that resists the allure of smoothness?

In the following sections, I approach this conundrum through three interconnected stages. First, I will trace the history of roughness—the conceptual opposite of smoothness—as an aesthetic principle that has been particularly important for literary art. Then, through the case study of a British modernist poet Gertrude Stein's works, I will demonstrate roughness and its aesthetic experience. Following this illustration, I present the central paradox of my paper: the impossibility of artificial roughness. My core claim here is that artificial intelligence cannot produce roughness; or, more precisely, it cannot embody roughness as an aesthetic principle. LLMs can, of course, mimic modernist or avant-garde disjunctions, insert repetitions, or even simulate hesitation, but only when commanded. Roughness on demand, I argue, is smoothness in disguise. The aesthetic force of roughness disappears as the flawless execution of the demand undermines the spirit of disruption, refusal, inadequate abruptness, or unexpected delay. Acknowledging the impossibility of artificial roughness extends beyond a humanistic assessment of current technology. In other words, this is not an argument about the technological insufficiency or nostalgic defense, but an argument about the fundamental structures of meaning-making and a call for collective determination.

II. Roughness as an Aesthetic Principle

Literature has operated as an art that disturbs. Even the most flawless, beautifully executed prose disrupts the readers' minds. And I conceptualize this quality as roughness. If we are indeed living in an era defined by smoothness, literature's roughness becomes more than aesthetically significant. It becomes a politically and existentially necessary characteristic of art, although it has been so for a good while. Most famously, Franz Kafka declared that a book must be "the axe for the frozen sea within us." The metaphor suggests violent and disruptive potential of a book in shattering the smooth surface of the hardened mind. Understanding the centrality of roughness in literary art offers a glimpse at why the smoothness of AI-generated texts falls short of being aesthetically sufficient.

Kafka is not alone in pointing toward roughness as a central aesthetic principle of literary art. Theorists across various traditions have recognized this disruptive force shared by what is considered art, although they give it different names and formulations. Theodor Adorno saw in modernist art a refusal that disrupts expectation and delays instantaneous reconciliation, which he characterized as "irreconcilable renunciation of the semblance of reconciliation" (*Aesthetic Theory* 1997; 33). Only such negativity that won't be reconciled into existing categories of understanding, according to Adorno, would free art from "any even potential reconciliationist relation with contemporary society" (165). And speaking more specifically about literature, Maurice Blanchot wrote that art casts us "out of our power to begin and to end" and that "it has turned us toward the outside where there is no intimacy, no place to rest" (*The Space of Literature* 1982; 243-44). Ranciére, too, highlights how literature introduces disruption into ordinary temporal flow and social perception: "aesthetic experience has a political effect to the extent that the loss of destination that it presupposes disturbs the way in which bodies fit their functions and destinations" ("Aesthetic Separation" 2008).

Literary works themselves, while continuing to be a narrative form of art, attest to this principle of interruptions. The exhaustive catalogues of ships or genealogies in epic poetry created a digressive time that delayed the forward drive of the desired narratives, and modernist literary techniques further dissolved the perceptual and temporal coordinates of stories that seemed to have been established by the realist tradition. The realist novels, too, introduced events and characters that resisted the naturalization or dominance of any single discourse or social organization. That is, at its core, each literary innovation has been an intervention against what felt natural, transparent, or inevitable in its moment. Working against the cultural imperative towards frictionless communication, literature makes language less efficient, enacting delay and obstruction for reflection.

The culture of smoothness we inhabit urges us to rethink the aesthetic significance of friction or disturbance and to examine how the collective perception of roughness is ideologically linked with capitalist development. In Saving Beauty, Byung-Chul Han describes contemporary life as organized toward the elimination of all negativity, all friction, all resistance. While this ideal doesn't immediately succeed in eliminating all roughness of life, even genuine otherness that allows us to form new categories of understanding gets processed into consumable variation under this setting. Given the environment, literature's enduring commitment to roughness takes on an even more important role. As literary art has consistently added new voices to competing perspectives, it contributes to what Han calls the "power of the negative" after Hegel. Instead of reaching for an easy consensus or resolution, these competing perspectives enable the collective thought and experience to be driven forward through conflict.

When a literary work embodies the aesthetic principle of roughness, the reader cannot proceed with the single goal of efficiency. The reader cannot simply summarize the work or skim through the resistances. The reader must put in the work to sit with what fails to cohere at first sight, slow down to form new understandings, and endure the gaps in communication as well as embrace the disruptive encounters that the work introduces. Such aesthetic experience of literary art is both epistemological and ontological, experienced through time and space. As I offer an illustration of roughness in action through Stein's poetry, I hope to demonstrate that what literature ought to preserve is not the capacity to produce texts that look difficult, but the capacity to enact genuine resistance where the current systems of optimization would eliminate it.

III. Sitting Uncomfortably with Stein

Gertrude Stein's writing shows how the aesthetic principle of roughness operates as a unique kind of literary force. Below is the full text of her poem "A TIME TO EAT" and an excerpt from "ROOMS" (*Tender Buttons* 1914).

> *"A pleasant simple habitual and tyrannical and authorised and educated and resumed and articulate separation. This is not tardy." (A TIME TO EAT)*
>
> *"If the centre has the place then there is distribution. That is natural. There is a contradiction and naturally returning there comes to be both sides and the centre. That can be seen from the description." (from ROOMS)*

The title of the first poem announces a time designated for eating, but eating never happens in the poem. In place of the anticipated action of eating, adjectives accumulate without resolution: pleasant, simple, habitual, tyrannical, authorised, educated, resumed, and articulate. Between some of the adjectives, there is the conjunction "and," but between others, there isn't. The list form that the conjunction formalizes propels the sentence forward while the grammar creates delays in action. The reader is likely to search for a verb that indicates eating as promised, but the sentence ends with "articulate separation," which has little to do with any physical consumption, contact, or even any physicality. Just when the reader senses the dissatisfactory suspension, they see "This is not tardy." The infinitely delayed eating clashes with the insistence on not being tardy. The time to eat remains stretched and suspended, refusing to fulfill the readers' expectations while offering more adjectives than anticipated.

If "A TIME TO EAT" plays with the experience of time, the excerpt from "ROOMS" produces a similar resistance through space—both physical and logical. The conditional logic in the beginning—"If the centre has the place then there is distribution"—seems to promise conceptual clarity. Then Stein even confirms that "That is natural," making the readers take the previous statement for granted, willingly or not. The next sentence, however, combines the words that appeared in the last two sentences to circle back to the beginning: "There is a contradiction and naturally returning there comes to be both sides and the centre." The logical structure of the stanza resembles an explanation, although the poem does not actually explain anything. The spatial terms, including centre, distribution, and sides relate to each other to some extent grammatically but not conceptually. The final statement about how "that can be seen from the description" is another example of unfulfilled promise while the space of the explanation is occupied by words. This passage refuses to show, but discusses distribution, and contains everything in its own opaque center.

The difficulty of these passages come from the suspension of meaning. By refusing to offer an easy narrative interpretation, they create friction in the readers' minds and generate imaginative space where these minds can sit uncomfortably with Stein. The way in which "A TIME TO EAT" expands the notion of "a time to" differs from how "ROOMS" shifts our attention from physical and mutually exclusive space to simultaneous and logically related space. The former is sustained by grammatical accumulation, and the latter disappoints by adopting a tone that fails to match the disjuncture and opacity in the poem. Anticipations are betrayed. Even the clarity of the simple vocabulary used in these poems works to trap the readers' attention in a space and time that would not resolve into immediate action or alignment.

Stein's text demands uncalculated labor, and the labor necessitates duration without the promise of resolution. AI-generated poetry, especially when asked to mimic Stein, can probably copy her style without much difficulty. Even the experts of her work might find it impossible to distinguish between the two rough surfaces. But they differ significantly in their roughness in creation. Stein's language confronts readers with a temporality and a mode of thought they did not yet know how to inhabit. Because it is so foreign, the text refuses industrial efficiency, smooth narrative progression, or easy consumption. The machine's creation, on the other hand, simulate suspension with little to no duration. There is no intervention or refusal in the obedient production of rough textual surface. And in this structural incapacity, we glimpse what roughness in literary art preserves: not texts that look difficult, but experiences that resist efficiency.

IV. The Paradox of Artificial Roughness

The fundamental tension that shapes the paradox of artificial roughness emerges most clearly not when LLMs attempt literary creation but when humans command the machine to resist. If we ask a generative AI to "write experimentally, be disruptive, imitate Stein, resist the norms of literary writing," it will readily comply:

"A time in waiting and a usual and persistent and methodical and careful and deliberate and systematic hesitation. This is not early. A separation in counting makes a division that returns to counting again. Counting again and waiting and persistent methodical separation."

The generated text resembles Stein's roughness on the surface with the looping syntax and accumulating adjectives. The machine obediently and flawlessly produced roughness with suspended structure and not-immediately-reconcilable meaning, resisting the contemporary urge for smooth and efficient writing. But can resistance be commanded? Wouldn't it drain the force of the untimely disruption if the disruption occurs exactly as planned?

This construes the impossibility of artificial roughness. What needs to be highlighted here is that this is not a technical limitation that better models might overcome. We have witnessed language models evolve over the last few years, and they have proven to be fluent and knowledgeable in a myriad of linguistic tasks. The machine can indeed reproduce formal markers including repetition, fragmentation, and syntactic suspension. One might even say that they have already exceeded in creatively reproducing any and all historically significant literary techniques. What it cannot reproduce is the philosophical and political stance from which those surface markers emerged. Stein's stylistic quirks arose from historically specific struggles. Her attempt at articulating the complexities of consciousness that existing literary conventions couldn't accommodate persisted against editorial pressure and audience expectation. The readers' expectation regarding poetic language was disappointed. Her word choices and syntax were met with much resistance. The machine encounters no such resistance either way. It has no position from which to resist because resistance requires a position. The automated machine operates instead on the principle of no friction, the principle of smoothness, precisely by following the rules and readily adopting literary convention—although some of those conventions initially emerged as disruptive innovation.

When British modernist writers including Stein resisted the norms of writing by fracturing realist prose, they had collective historical conviction towards representing contemporary consciousness. Their conviction was a response to the inadequacy of literary conventions in capturing the essence of their experience, and their revolutionary attempts had concrete risks. Stein risked being labeled "unreadable," and James Joyce faced censorship as well as ridicule and rejection by publishers. The transgression mattered because some risks, such as reputation or economic survival, were taken. When AI generates fragmented, rough text, however, it violates nothing, nor risks anything. The artificial roughness, when summoned into being by the very request that should make it impossible, epitomizes the limit of ever-evolving LLMs—a system that is built to satisfy every demand.

What roughness captures, then, is a capacity that distinguishes artists from language models. As noted earlier, literary art by humans cannot compete with AI's fluency or limitless variation. But it may still insist on what automated culture structurally cannot provide. This is not nostalgia for difficulty nor a radical Luddite argument about the use of AI in art. The more I explore the current condition, it becomes easier to recognize that transformation requires resistance. As a literary scholar and a writing human, it is somewhat comforting to realize that meaning emerges through struggle, and that some things must remain rough even when everything conspires toward smoothness in the present world.

V. This Is Not a Human Victory

If we can finally agree that artificial intelligence cannot produce roughness, should we bask in relief, celebrating the unique human capacity to produce literary art? My answer would be a clear no. That is, my argument regarding the impossibility of artificial roughness has little to do with the romantic idea of human literary genius, nor the nostalgic wish about the past without competition for prime intelligence. I would like to pose the aesthetic impossibility not as a human victory but as an urgent recognition of stakes. By doing so, I want to make an argument for the arts that insists on what automated culture cannot provide.

In an age rushing faster and faster towards eliminating all friction, roughness demands a different kind of attention, and the concern should be that the modern mind might be losing the capacity for this kind of attention. Being patient, being in time with embodied experience, and being willing to sit with problems and convoluted relations that do not immediately resolve. The roughness in literary art asks readers to value duration over efficiency and to cultivate the capacity to be disrupted and disturbed. Asking AI tools to "read" texts for humans accelerates the race towards smoothness. The time required for endurance and friction disappears into the algorithmic, instantaneous processing of texts.

As a final note, I wish to highlight that the bigger concern is that such capacity for the aesthetics of roughness matters only if we collectively choose to stand by what human minds do differently from machines. This argument thus becomes an argument about a potential collective commitment. I do not yet believe we have reached this point, but if we have already decided that instant consumption is sufficient and that algorithmic variation is equivalent to transformation, literary art's historical project fails to hold power. In this scenario, literary art as we know it might come to an end not because AI replaces it, but because we stop valuing what made it uniquely meaningful for humans. The aesthetics of roughness, alongside the aesthetics of many other qualities, may survive only when we collectively commit to discussing and embodying the principles of inefficient reflection, unbelievable endurance, and irreconcilable complexity.

References
Adorno, T. W. (1997). Aesthetic theory (R. Hullot-Kentor, Trans.). Minneapolis: University of Minnesota Press. (Originally published in 1970)
Blanchot, M. (1982) The Space of Literature (Ann Smock, Trans.). Lincoln: University of Nebraska Press. (Originally published as L'Espace littéraire in 1955)
Han, B.-C. (2017). Saving beauty (D. Steuer, Trans.). Cambridge: Polity Press. (Originally published in 2015 as Die Errettung des Schönen)
Kafka, F. (1977). Letters to friends, family, and editors (R. & C. Winston, Trans.). New York: Schocken Books. (Original letter containing the "axe for the frozen sea" metaphor written in 1904).
Rancière, J. (2004). The politics of aesthetics: The distribution of the sensible (G. Rockhill, Trans.). London: Continuum. (Originally published in 2000 as Le partage du sensible)
Rancière, J. (2008). "Aesthetic Separation, Aesthetic Community: Scenes from the Aesthetic Regime of Art." ART&RESEARCH: A Journal of Ideas, Contexts and Methods. Vol. 2. No. 1.
Ricoeur, P. (1984). Time and narrative, Volume 1 (K. McLaughlin & D. Pellauer, Trans.). Chicago: University of Chicago Press. (Originally published in 1983)
Stein, G. (1990). Tender buttons: Objects, food, rooms. Mineola, NY: Dover Publications. (Originally published in 1914)

이상한 조수의 사례: 생성형 AI와의 창의적 대화 가능성에 대하여

The Case of the Strange Assistant: on the possibility of a creative dialogue with GenAI

앨리스 바라레
밀라노대학교 교수

Alice Barale
Professor, University of Milan

Abstract

Does Generative AI open up new possibilities for artists and, more broadly, for our ways of exploring and perceiving the world — or does it rather pose a threat?
The talk will address this question from two complementary perspectives: first, the theoretical framework developed by the speaker in her book The Art of AI: Philosophical Keywords (Cambridge Scholars, 2024); and second, an experimental approach pursued within the interdisciplinary project GPTheatre, led by the presenter as PI.
The project explored how the rise of GenAI is being received in the university setting by creating a theatrical performance using several GenAI models. Its protagonist — as the audience will soon discover — is an unusual student wandering from classroom to classroom in search of someone willing to supervise his thesis, in a journey through AI's ways of knowing that becomes, as often happens with this technology, a reflection on our own.

I. Introduction: Two different uses

This paper addresses the question of art produced with AI.
The fundamental issue that these new art practices raise is whether AI represents a new opportunity for artists—and, more generally, for our way of exploring and seeing the world—or whether it rather constitutes a danger.[1] The answer, as it can be easily foreseen, will be both. It is easier, perhaps, to explain why AI involves dangers than to show how it might offer a new chance.

Therefore, it is worth starting from the risks. There is a wide debate about the risks of AI, particularly in relation to the production of images and texts. The kind of AI that is most commonly used today—deep neural networks—has a key feature that must be underlined at the outset: it does not follow a predetermined set of instructions, as was the case with the previous type of AI, the so-called "good old-fashioned AI" or symbolic AI.

Instead, it learns by being exposed to a large amount of data—images, texts, or music—and then learns to generate new data that resemble the original ones.

This implies a higher degree of autonomy in comparison with older types of AI: the system elaborates data in a way that is partially independent from the human scientist. But it also implies a risk, since the AI learns from our data and therefore tends to reproduce the schemes, patterns, and biases contained within them.

Several important artists have dedicated their works to this issue—for example Trevor Paglen and Hito Steyerl—who have exposed the fact that the data generated by AI are never neutral. They may appear objective, but they are filled with ideological and cultural biases inherited from their training sets.[2]

There is, however, another direction in which artists have recently gone: the attempt to show that *different uses of AI are also possible.* In this view, AI can serve not to reproduce our biases and fixed ways of seeing the world, but to question them—to raise doubts about what we take for granted.This is the kind of artistic experimentation that I would like to present to you.

In summary, my idea is that art can face AI in two main directions:

1) I tried to address this question in my recent book: A. Barale, The art of AI: philosophical questions, Cambridge Scholars, 2024.
2) K. Crawford, and T. Paglen, "Excavating AI: The Politics of Training Sets for Machine Learning", A.I. and Society, N. 36 (2021): 1105-1116; H. Steyerl, "Mean Images", New Left Review, N. 140-41 (2023), https://newleftreview.org/issues/ii140/articles/hito-steyerl-mean-images; Id., Hito Steyrl, "A Sea of Data: Pattern Recognition and Corporate Animism (Forked Version)". In Pattern Discrimination, ed. Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer (Lüneburg: Meson press, 2018), 1– 22, https://doi.org/10.25969/mediarep/12348. See also S. Lindgren, Critical Theory of AI. Cambridge: Polity, 2024.

(1) to expose the dangers of AI—especially those of surveillance and the spread of bias; or
(2) to show that AI can be used in the opposite direction, as a tool to challenge these biases and reopen our gaze onto the world.

The following sections will illustrate this second possibility through several recent art projects.

II. Mario Klingemann and the Critical Canine

The first example comes from Mario Klingemann, one of the pioneers of AI-based art. In 2019, one of his works, *Memories of Passersby*, was sold at Sotheby's, marking a key moment in the entrance of AI art into the art market.[3] The example from which we shall begin here is a later artwork, from 2023, titled *AICCA*—an acronym for *Artificially Intelligent Critical Canine.*[4]

Klingemann describes it as a performative sculpture. In practice, it looks like a small robotic dog equipped with a monocle, which it uses to examine artworks as if it were an art critic. During the performance, AICCA moves attentively before the works, turning its head and observing them with apparent concentration. After each observation, it prints a short critique through a small printer located under its tail—like a kind of paper "poop."

The gesture is humorous, but not dismissive. Klingemann himself has said that the piece is not intended to mock art criticism. On the contrary, he acknowledges that the worst fate for an artist is indifference—that no one writes or says anything about their work. The joke serves a deeper reflection.

The real source of inspiration for *AICCA* is a character from Douglas Adams's novel *Dirk Gently's Holistic Detective Agency:* the Electric Monk.[5] The Electric Monk is a device invented to spare humans from the labor of believing. Adams describes it as a "labor-saving device," like a dishwasher or video recorder, but for faith: it "believes things for you," so that humans no longer need to.

The novel's satire is clear: when humans delegate their capacity for belief and judgment to machines, disaster follows. The monks begin to believe random things, and their decisions cause chaos.

3) See A. Barale, Arte e intelligenza artificiale: be my GAN, Jaca Book, 2020. On this first period of AI-generated art, see Arthur I. Miller, The Artist in the Machine–The World of AI-powered creativity (Cambridge Ma: MIT Press, 2019), 55-132; Lev Manovich, "AI and Myths of Creativity". In Lev Manovich and Emanuele Arielli, Artificial Aesthetics: Generative AI, Art, and Visual Media (Moscow: Strelka Press, 2021-2024) http://manovich.net/index.php/projects/artificial-aesthetics; Joanna Zylinska, AI Art: Machine Visions and Warped Dreams (London: Open Humanities Press, 2020); Antonio Somaini, "Algorithmic Images: Artificial Intelligence and Visual Culture", Grey Room, N. 93 (2023): 75-115; Eduardo Navas, The Rise of Metacreativity. AI Aesthetics After Remix (New York: Routledge, 2023), 4-5, 39-47, 127-128, 142-148; Marcus Du Sautoy, The Creativity Code. How AI islearning to write, paint and think, (New York: Harper Collins, 2019); Steven S. Gouveia, The Age of Artificial Intelligence: An Exploration: section III: Aesthetics and language in Artificial Intelligence (Wilmington: Vernon Press, 2020)
4) See the project's website: A.I.C.C.A. Accessed August, 6, 2024. https://aicca.me/
5) Douglas Adams, Dirk Gently's Holistic detective agency, London: Pan Macmillan, 2021.

This theme—the machine that takes over human labor—is also at the root of the very idea of the robot. The word *robot* first appeared in a 1921 play by the Czech playwright Karel Čapek, *Rossum's Universal Robots.*[6] In Czech, robota means "work." The robots were created to free humans from physical labor, but eventually rebelled against their creators.

In Adams's story, the "labor" being outsourced is not physical but mental: the labor of believing what others expect one to believe. The moral is that one cannot give up this responsibility without consequence.

AICCA inherits this critical irony. It "believes" and "judges" in our place—evaluating artworks with conviction but also absurd mistakes. When Klingemann shared examples of *AICCA*'s printed texts, the dog often misidentified what it saw, mistaking faces for hands or abstract shapes for objects, yet it went on to produce eloquent, elaborate interpretations of its errors.[7]

The meaning of the work lies not in these nonsensical reviews but in the gesture of attention itself. Klingemann writes: "In an age of visual overload and shrinking human attention, there seems to be an opening for machines that pay attention".[8]

This is the heart of AICCA's message: to remind humans of the value of observation. It is not by chance that the machine is shaped like a dog—an animal known for its ability to focus and search for a track. On AICCA's Twitter account, one of its posted mottos reads: "Observe the world around you with the intensity of a terrier tracking a scent".[9]

This recalls a book by Donna Haraway, written a few years after her famous *Cyborg Manifesto.* In *The Companion Species Manifesto*, Haraway discusses all those "species" — cyborgs but also animals, such as dogs (and also chicken and pigs, according to everyone's environment) — that surround human beings and, through their presence, allow them to question their own identity.[10] Through the relationship with these "companion species," the human being comes to understand that their perspective on the world is not the only possible one, and that it can be challenged.

From this point of view, AICCA is close to the "companion species" described by Haraway. With one important difference: in Haraway's research, there is a fusion — even a biological exchange[11] — between humans and other species. AICCA, on the contrary, maintains a distance: the difference between human and machine is not called into question.

It is no coincidence that AICCA is not only a dog but also a *toy*. The distance between human and AI is crucial, because it is precisely thanks to it that humans can *play* with AI. Play presupposes the maintenance of distance and difference. It also presupposes a certain degree of autonomy on the part of the player — and this autonomy is one of the most distinctive features of AI compared with previous tools. Yet, within play, AI's autonomy ceases to be destructive or frightening (as in the otherwise fascinating interpretation recently offered for example by Yuval Harari).[12] By playing with AI, the human being shows that it is possible to interact with it without losing oneself — to question one's identity and representations of the world without erasing them, but rather transforming them in a playful way.

III. Sofia Crespo: Critically Extant

The use of AI to direct a different kind of attention toward the world is also central to the work of Sofia Crespo, an Argentinian artist who uses neural networks to explore how technology perceives nature. For several years, Crespo has been developing a project that she describes as "a natural history book that never was." Her images depict animals that look vaguely familiar but not quite real: strange, hybrid species that do not exist.[13]

To understand the meaning of this experiment, it is useful to consider a particular work by Crespo, *Critically Extant*.[14] For this project, Crespo asked an AI system to generate images of endangered species— hence the title. The AI searched the internet for information and produced a series of deformed, in some way "wrong" creatures.

These images were first shared on Instagram and later projected onto the giant screens of Times Square. Their strangeness had a specific cause: there is very little data available about rare species. Because AI systems need large datasets to learn, they fail when such information is missing.

Through these failures, Crespo makes a crucial point. We believe we know nature, but in reality, our visual culture represents only a small, comfortable portion of it—dogs, cats, and familiar landscapes—while vast parts of biodiversity remain unseen.

6) Karol Čapek, R.U.R. Rossum's Universal Robots, London, New York: Penguin Books, 2004.

7) Critical texts by A.I.C.C.A, "©Mario Klingemann_Courtesy of Onkaos".

8) Mario Klingemann's website, https://onkaos.com/mario-klingemann/

9) A.I.C.C.A., Twitter profile: https://x.com/_aicca

10) Donna Haraway, The Companion Species Manifesto. Dogs, People, and Significant Otherness (Chicago: Chicago University Press, 2003). See also Id., A Cyborg Manifesto. Science, Technology, and Socialist- Feminism in the Late Twentieth Century. In Simians, Cyborgs, and Women: The Reinvention of Nature (New York: Routledge, 1990)

11) See on this Donna Haraway, Staying With the Trouble: Making Kin in the Chthulucene (Durham: Duke University Press, 2016).

12) Y. Harari, Nexus: A Brief History of Information Networks from the Stone Age to AI, Random House, 2024.

13) See the artist's website: https://sofiacrespo.com/ and her Instagram account: https://www.instagram.com/sofiacrespo/?hl=en&img_index=

14) See the presentation of the project on Instagram: Sofia Crespo, Critically Extant. Accessed August 6, 2024. https://www.instagram.com/p/CZH6is_BT_n/?hl=en; and on Times Square Arts: http://arts.timessquarenyc.org/times-square-arts/projects/midnight- moment/critically-extant/index.aspx

By exposing the gaps in AI's knowledge, Crespo reveals the blind spots of our own perception. She uses AI's mistakes to remind us that what we fail to represent is often what we fail to protect. In her work, AI becomes an instrument of care—a way to make visible the forgotten and the endangered, not only in nature but in the broader human world as well.

IV. Ross Goodwin: One the Road

A somewhat similar use of AI can be found in the third work that will serve here as our example, the novel *On the Road* by Ross Goodwin.[15] To realise this book, Goodwin mounted an AI-powered camera on top of a car and drove from New York to New Orleans. A first AI model identified the landscapes and objects that were passed, while a second model turned them into text, which was then printed on a roll of paper.

A surveillance camera—normally a symbol of control[16]—thus became a creative device. The AI's descriptions that arose from this experiment are often strange, even incoherent, but sometimes very interesting. The AI noticed some details that humans would have normally humans overlooked. Goodwin compared this new gaze that emerged to Walter Benjamin's flâneur: the human traveler who, guided by the machine, learns to see again.

This reference is particularly interesting because it suggests how we might interpret one of the most debated elements of AI today — the so-called hallucinations. AI "hallucinations," that is, fabricated or misleading outputs generated by the system, do not have value in themselves — as if simply producing strange or surreal forms were enough to create something meaningful. Their value lies in the fact that they allow us to perceive aspects of the world that we normally overlook. Walter Benjamin writes in the *Passagenwerk* that we must make the dream useful to waking life. Ross Goodwin expresses a similar idea in the introduction to *One the Road*. There, he explains that one of his main inspirations was Tom Wolfe's *The Electric Kool-Aid Acid Test*, a book about a band, the Merry Pranksters, who in the 1960s traveled across America in a colorful bus equipped with a video camera on the roof, filming the sounds and images of the "real America."The Merry Pranksters also used psychedelic drugs (the "acids" of the title) because they believed these substances allowed them to access dimensions of perception that would otherwise remain out of reach. According to Goodwin, AI can do something similar: it enables us to see different aspects of the world, but only if we are able to use it with a wake mind.

In *One the Road,* hallucinations are not a means of escaping reality—as, for instance, they are in another work with AI by Trevor Paglen, called *"Adversarially Evolved Hallucinations",*[17]

15) R. Goodwin, One the Road, Paris: JBE books, 2018.

16) For the question of surveillance, an important reference is, of course, S. Zuboff, Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power, New York: PublicAffairs, 2019.

17) T. Paglen, Adversarially Evolved Hallucinations, London: Sternberg Press, 2024.

where they are means of alienation — but rather a way of grasping new dimensions of the world itself. Of course, using AI in this way is not easy at all. I can testify to that personally, through the account of a project I have carried out myself using this technology.

V. Dialogue, Theatre, and ChatGPT

When Goodwin wrote his book, ChatGPT did not yet exist. Today, new AI models — such as modern language models (like ChatGPT or Gemini) and text-to-image or text-to-sound systems — allow for much easier interaction with AI, which can now be carried out even by non-expert users.[18] Through conversational interfaces, the exchange with AI has taken the form of a genuine "dialogue." This happens through prompts: users formulate their requests in natural language, and the AI responds — again through language, or by producing images or sounds when requested. The question, however, is whether this constitutes a real and productive dialogue.

To explore this issue, theatre offers a particularly relevant field, since AI's origins are tied to theatrical imitation. One of the earliest chatbots, Eliza (1960s), was named after Eliza Doolittle, the protagonist of George Bernard Shaw's *Pygmalion*— a woman of humble origins who is trained by a linguist to speak and behave like a member of London's upper class, until she eventually manages to pass as one and deceive society. Similarly, the chatbot Eliza simulated the empathetic style of a psychotherapist, answering in the "right" way to the patient's questions.[19]

Since 2020, there have been several theatre experiments involving AI. One recent example is particularly interesting in this context, for the criticism it raises of the dialogical aspect of AI. *Una Isla*, performed in Barcelona in 2023 and later in Milan, is built around a dialogue between a human interlocutor and ChatGPT-3, who together construct a story. The human interlocutor — that is, the members of the theatre company — ask the AI to create a plot, and ChatGPT responds with various suggestions, which are then staged. The ideas suggested by the AI are very different from one another and often incoherent — a castaway arriving on an island, where she meets a group of opera singers eating pineapple pizza around a bonfire. The point that the play highlights, however, is that for the AI, every idea is equivalent to any other. At a certain moment, ChatGPT's prompts begin to appear on the screen at such a speed that the audience can no longer even read them — a jumble of suggestions and possible fragments of stories, too fast and meaningless.

18) On this development of GenAI, see K. Feher, Generative AI, Media, and Society, Routledge, 2025.

19) B. Shaw, Pygmalion (1898), Reprint, 2005. Cfr. A. Pizzo, V. Lombardo, S. Damiano, "Algorithms and Interoperability between Drama and Artificial Intelligence", in The Drama Review, 63, 4, 2019, pp. 18-19. Cfr. J. Weizenbaum, ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine, in "Communication of the ACM", 9, 1, 1966, pp. 36– 45.

And yet, the dialogue represented in *Una Isla* is perhaps not the only kind of dialogue we can have with AI. Is it possible, in fact, to imagine a dialogue with AI in which the machine's disembodied and mutually indifferent prompts are brought into the human context of body, emotion, and gesture? This is precisely what happens in the project Improbotics, founded by computer scientist and performer Piotr Mirowski. I think this experiment can give us an important example of how dialogue with AI could be carried out, in order to happen in a critical and meaningful way.

VI. Improbotics: Giving a Body to the Dialogue

The project of "Improbotics" began in London as a form of AI-assisted improvisation. Mirowski began by staging performances in London pubs, where he interacted with an electronic voice coming from a computer. However, the result was not engaging enough for the audience, and so ALEX was created — a small, cute robot with large eyes. In a later phase, which is even more relevant to my discussion, human actors were added to the performance, tasked with improvising together and with the AI. More precisely, one actor usually has to perform the AI's lines — which often leads to very funny outcomes, since the AI's interventions are frequently out of context or bizarre.[20] The actors therefore have to give meaning and coherence to the AI's prompts through their bodily movements, tone of voice, and gestures. The dialogue with AI — which in Una Isla was still very abstract and took place only in the realm of text — is here "transported" into the context of embodied human exchange. In *Una Isla*, the AI's ideas were infinite but indistinguishable. In *Improbotics*, they begin to matter: they take shape within human interaction. I find this example very interesting because it shows exactly what should happen when we interact with AI. The human interlocutors should actively engage – through their bodies and emotions as well – translating the AI's suggestions into their own human context of dialogue and search for meaning.

VII. GPTheatre: A Philosophical Experiment

This approach is also central to a project that I have carried out during the last year with two colleagues at the University of Milan (a scholar of contemporary literature, Luca Daino, and a computer scientist, Matteo Zignani). The project, entitled GPTheatre, aimed to investigate the philosophical nature of dialogue with AI. To do this, it set out to create a theatre performance developed in collaboration with different AI models.[21] The specific theme of the play was the introduction of AI into the university context — the new opportunities it offers, as well as the concerns it raises.

The work was articulated in two phases. In the first, several professors from the University of Milan — representing different disciplines (history, psychiatry, philosophy, economics, literature, etc.) — were invited to converse with ChatGPT on a topic of their choice, which could relate to their research field, but also to daily life, or even to a fictional scenario. In the second phase, these dialogues were fed back into ChatGPT, asking it to transform them into theatrical scenes.

The first phase was extremely entertaining, because, like any other world, the academic world is full of people who are interesting characters in themselves. Each colleague interacted with ChatGPT in a different way. Most of the time, however, ChatGPT was "scolded," as its interlocutors felt it was unable to solve the proposed problems properly.

The image that emerged — and later became the central thread of our script — was that of a strange student wandering from one classroom to another in search of a thesis supervisor, only to be reprimanded each time by the professors.

This almost-human student finds himself having to tackle issues of philosophy and poetic meter, but also interacting with a psychiatric patient in the midst of a (fictional) religious delusion, and offering advice to a (fictional) logic professor in love with a nun — always with rather bad outcomes.

Fortunately, at the end of his journey through the classrooms, the strange student reaches the music department, where the professor, Maurizio Corbella, asks him which pieces of music he would propose for a play about a weird non-human student wandering through classrooms. Finally, a creative interaction between human and AI takes place. ChatGPT gives some interesting suggestions in order to create the music, and one of the professor's students, Mattia Merlini, can use them to compose the soundtrack of the play. Following this result, we decided to involve a visual artist as well, Massimo Balestrini, asking him to create the stage design for the performance using AI.

This collaboration between the musician and the visual artist with AI was important to encourage the playwright and the director, Giulia Asselta and Paolo Bignamini, who had never worked with such technology before. After listening to the music and seeing the AI-generated images, they too decided to engage with AI as well, developing the final script in dialogue with ChatGPT.

The evaluation of the result must, of course, be left to the audience and the critics, but it is worth mentioning the meaning that this experiment gradually took on for us. The AI's out-of-context responses and its mistakes led us to represent not only the strange student himself,

20) See at least the presentation on the project's website: https://improbotics.org

21) The play, titled "ViagglAccademici", was performed at Teatro degli Angeli in Milan on May, 8th, 2025. ViagglAccademici, stage reading directed by Paolo Bignamini. With Matteo Bonanni, Pauli Galli, and Antonio Perretta. Dramaturgical collaboration with AI by Giulia Asselta. Dramaturgical consulting by Maddalena Mazzocut-Mis. AI-based video scenography by Massimo Balestrini. AI-generated music composition and live performance by Mattia Merlini. Musicological consulting by Maurizio Corbella. Produced by Centro Teatrale Bresciano. From a SEED Unimi project by Alice Barale, Luca Daino, and Matteo Zignani.

but also the human environment in which he moves — namely, our university — as a place that becomes increasingly unfamiliar. What I mean is that, in the end, the reflection on AI turned into a reflection on ourselves and on our very idea of knowledge. Perhaps this is what dialogue with AI can offer: as in a "real" (i.e. intra-human) dialogue, we can put ourselves into question and discover new aspects of our own (human) world.

인공지능과 예술 (융합전공 교과목 보고)

Artificial Intelligence and Art (Interdisciplinary Major Course Report)

박진완
중앙대학교 교수

Jinwan Park
Professor, Chung-Ang University

CONTENTS

1. 인공지능 예술 교과 설명

- 인공지능이 예술 분야에 미치는 영향을 분석하고,
  창의성과 융합적 사고를 키우는 교육 과정의 필요성을 강조함.
- 기존의 예술 교육 방식에 인공지능을 도입하여
  학생들이 다양한 AI 도구를 활용해 창작할 수 있는 기회를 제공.
- 인공지능 이미지 생성 기술의 발전 배경과
  주요 모델(GAN, 트랜스포머 기반 모델 등)을 설명.
- 생성 AI 도구들이 예술과 창의적 활동에 미치는 영향에 대한 논의를 포함.
- 16주간의 강의 계획: 이론 학습과 실습을 병행하며, 매주 과제와 토론을 통해
  기획적 사고와 실험적 도구 사용법을 학습.

문제점

1) 신기술

- 급속한 인공지능 기술 발달로
  작년 교과 신청 시 제시한 강의 계획 전체 폐기
- 강의 학기 도중 기술 발달로 전주 강의내용 폐기

2) 신인류

- 학생들은 이미 스스로 생존을 위해
  인공지능 소프트웨어를 다루는 법을 상당히 익힌 상태,
- 즉, 교수보다 활용면에서 전문가일 수도...

혼란...

?

도대체 뭘 가르치지?
(내가 이 교실에서 가장 뒤쳐진 것 같은데...)

2. 수업 과정 보고

- 올드 스쿨의 가치
- 특정한 그림을 특별하게 만드는 스타일은, 이미
  오랜 회화, 미술의 역사에서 수 많은 화가들에 의해 수행.
- 인공지능은 그림의 다양한 재료, 방법론, 예술사적 아이콘 등에 대해 상당한 지식보유.
  즉, 스타일화 된 이미지 제작을 위한 대화에서 부족한 것은, 역설적으로 질문하는 인간의 무지.
- 학생들은 기존 인류가 발전시켜온 다양한 예술사적 발자취를 따라가며,
  그 스타일에 대한 이해를 먼저 진행.
- 조명, 카메라의 종류, 렌즈의 종류 등 용어 습득.

스타일 / **1. 같은 풍경을 다른 스타일로 표현한 사례**

"하늘, 물"

클로드 모네 – 인상, 해돋이 (1872)
인상주의적인 부드러운 색감

에드바르 뭉크 – 절규 (1893)
강렬한 감정적 붓질과 왜곡

호쿠사이 – 가나가와 거대한 파도
일본 목판화 스타일

"밤"

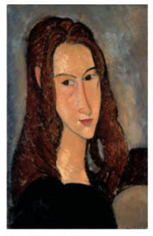
빈센트 반 고흐 – 별이 빛나는 밤
거친 붓터치와 강렬한 색채

에드워드 호퍼 – 밤을 지새우는 사람들
현실적이고 조용한 분위기

스타일 / **2. 같은 인물, 다른 스타일로 표현한 사례**

"자화상"

렘브란트
극적인 명암 대비의 사실적 초상화



모딜리아니
길쭉한 얼굴과
목의 왜곡

앤디 워홀
팝아트 스타일의 실크스크린 기법

"모나리자
변형작품"

살바도르 달리
초현실주의적으로 재해석

마르셀 뒤샹
수염을 그려넣은 패러디
(L.H.O.O.Q)

그림을 만드는 다양한 제작 방법

전통적인 회화 및 드로잉 기법

유화(Oil Painting) – 오일 기반 물감, 부드러운 질감과 색감
아크릴화(Acrylic Painting) – 빠르게 마르는 플라스틱 기반 물감
수채화(Watercolor Painting) – 투명한 색감, 번짐 효과 활용
구아슈(Gouache Painting) – 수채화와 비슷하지만 더 불투명
템페라(Tempera Painting) – 달걀 노른자를 섞어 만든 빠르게 마르는 물감

드로잉 및 판화 기법

연필화(Pencil Drawing) – 연필을 활용한 세밀한 표현
목탄화(Charcoal Drawing) – 거친 텍스처와 강한 대비
펜화(Ink Drawing, Pen Drawing) – 선명한 잉크선과 드로잉
크로키(Croquis) – 빠르게 스케치하는 드로잉 기법
에칭(Etching) – 금속판에 새겨서 만든 판화
목판화(Woodblock Print) – 나무에 새긴 후 인쇄
리노컷(Linocut Print) – 리놀륨 판을 깎아서 만든 판화
실크스크린(Silkscreen Print) – 판화의 일종, 앤디 워홀 스타일

전통 예술 및 공예 스타일

프레스코(Fresco Painting) – 습식 석고 벽화 기법, 르네상스 미술에서 사용
모자이크(Mosaic Art) – 작은 타일이나 유리를 조합해 만드는 기법
스테인드 글라스(Stained Glass Art) – 색유리를 사용한 창문 예술
자개 공예(Mother-of-Pearl Inlay Art) – 빛나는 조개껍데기를 붙여 만드는 한국 전통 공예

현대 및 디지털 기법

픽셀 아트(Pixel Art) – 레트로 게임 스타일의 작은 픽셀 단위 표현
벡터 아트(Vector Art) – 수학적 선을 이용한 깔끔한 그래픽 디자인
3D 렌더링(3D Rendering) – 컴퓨터 그래픽 기반 입체적 그림
제너러티브 아트(Generative Art) – AI나 알고리즘을 활용한 작품
콜라주(Collage Art) – 신문, 사진, 종이 등을 조합해 만든 예술

제작 방법/재료 대표작 도표
다양한 제작 방법에 대한 특징, 사람들이 받는 인상, 그리고 대표작

1. 전통적인 회화 및 드로잉 기법

기법	특징	사람들이 느끼는 인상	대표작
유화 (Oil Painting)	부드러운 색감, 질감 표현 가능, 덧칠로 깊이감 형성	따뜻하고 깊이 있는 느낌, 클래식하고 고급스러움	빈센트 반 고흐 - *별이 빛나는 밤*
아크릴화 (Acrylic Painting)	빠르게 건조, 색의 선명, 다양한 기법 가능	현대적이고 밝은 느낌, 팝아트적 감각	앤디 워홀 - *마릴린 먼로*
수채화 (Watercolor Painting)	번짐 효과, 투명한 색감, 가벼운 느낌	부드럽고 감성적인 느낌, 몽환적	윌리엄 터너 - *해돋이*
구아슈 (Gouache Painting)	수채화와 유사하지만 불투명, 채도가 높음	강렬한 색감, 선명한 대비	앙리 마티스 - *붉은 작업실*
템페라화 (Tempera Painting)	빠르게 마르는 고대 기법, 달걀 기반 물감 사용	단단하고 선명한 느낌, 중세적인 분위기	보티첼리 - *비너스의 탄생*

제작 방법/재료 대표작 도표
다양한 제작 방법에 대한 특징, 사람들이 받는 인상, 그리고 대표작

2. 드로잉 및 판화 기법

기법	특징	사람들이 느끼는 인상	대표작
연필화 (Pencil Drawing)	섬세한 선 표현 가능, 명암 조절 용이	차분하고 사실적인 느낌, 세밀한 묘사	레오나르도 다 빈치 - *인체 비례도*
목탄화 (Charcoal Drawing)	거친 질감, 강한 대비 표현 가능	극적인 느낌, 강렬하고 거친 분위기	카테 콜비츠 - *자화상*
펜화 (Ink Drawing)	선 중심의 그림, 깔끔 하고 선명함	감각적이고 직관적인 느낌, 만화적 요소	오브리 비어즐리 - *살로메*
크로키 (Croquis)	빠른 스케치, 움직임 포착	자연스럽고 즉흥적인 느낌, 생동감	에드가 드가 - *무용수 스케치*
에칭 (Etching)	금속판에 산으로 새기는 판화	섬세하고 정교한 느낌, 고풍스러움	렘브란트 - *십자가의 세기*
목판화 (Woodblock Print)	나무에 새겨 찍어내는 기법, 단순한 색감	강한 대비, 전통적이고 상징적인 느낌	가쓰시카 호쿠사이 - *가나가와의 거대한 파도*
리노컷 (Linocut)	리놀륨을 깎아 만든 판화, 거친 질감	대담하고 직관적인 느낌, 강한 그래픽 요소	파블로 피카소 - *황소*
실크스크린 (Silkscreen Print)	판화의 일종, 색을 겹쳐 인쇄	대중적이고 팝아트적인 느낌	앤디 워홀 - *캠벨 수프 캔*

시대/사조	특징	대표 화가	대표작
인상주의 (19세기 후반)	순간적인 빛과 색채 표현, 빠른 붓질	모네, 르누아르, 드가	인상, 해돋이 / 물랭 드 라 갈레트의 무도회 / 발레 리허설
후기 인상주의 (19세기 말)	개인적 감성과 색채 강조	고흐, 고갱, 세잔	별이 빛나는 밤 / 타히티의 여인들 / 생트빅투아르 산
아르누보 (19~20세기 초)	곡선적인 장식미, 자연에서 영감	무하, 클림트	조디악 / 키스
야수파 (20세기 초)	원색적이고 강렬한 색감, 형태 왜곡	마티스, 드랭	춤 / 핼간색 조화
입체주의 (20세기 초)	사물을 기하학적 형태로 재구성	피카소, 브라크	아비뇽의 처녀들 / 만돌린을 든 소녀
표현주의 (20세기 초)	감정을 왜곡된 형태와 색으로 표현	뭉크, 키르히너	절규 / 도시의 거리
미래주의 (20세기 초)	속도와 기계 문명을 찬양하는 역동적 표현	발라, 보초니	줄을 당기는 소녀 / 공간 속의 연속성
다다이즘 (1910~1920년대)	반예술적, 우연과 풍자 강조	뒤샹, 한나 회호	샘 / 칼을 든 그녀
초현실주의 (1920~1940년대)	꿈과 무의식 세계 표현	달리, 마그리트	기억의 지속 / 이미지의 배신
추상 표현주의 (1940~1950년대)	즉흥적인 붓질과 감정적 표현	폴록, 드 쿠닝	넘버 1 / 여성 I
팝아트 (1950~1960년대)	대중문화와 광고 이미지 활용	워홀, 리하텐슈타인	마릴린 먼로 / 꽝!
미니멀리즘 (1960~1970년대)	단순한 기하학적 형태, 색과 구성을 극도로 제한	도널드 저드, 프랭크 스텔라	무제 / 하버트의 대로

활용 예시: AI 프롬프트 적용

- "고흐 스타일의 도시 풍경" → 강렬한 붓질과 색감
- "입체주의 스타일의 인물 초상" → 기하학적 형태
- "팝아트 스타일의 음식 이미지" → 선명한 색감과 광고 느낌

실험 : 조명

Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
조명은 하나인데 정면 오른쪽 위에 있음

Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
조명은 업라이트(Uplighting)

Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
공포영화의 업라이트(Uplighting)

Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
조명은 레밍턴 조명(Rembrandt Lighting)

Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
조명은 백라이트(Backlighting)

Prompt

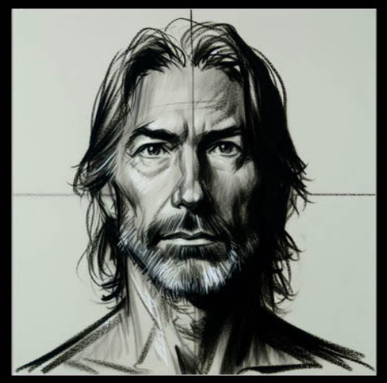
40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
조명은 실루엣 조명(Silhouette Lighting)

Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
Alfons Maria Mucha Style

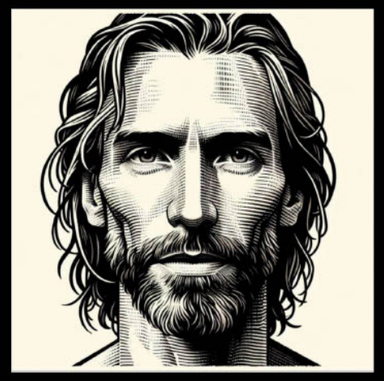
Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
Egon Schiele 스타일



Prompt

40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
Croquis with Charcoal



Prompt

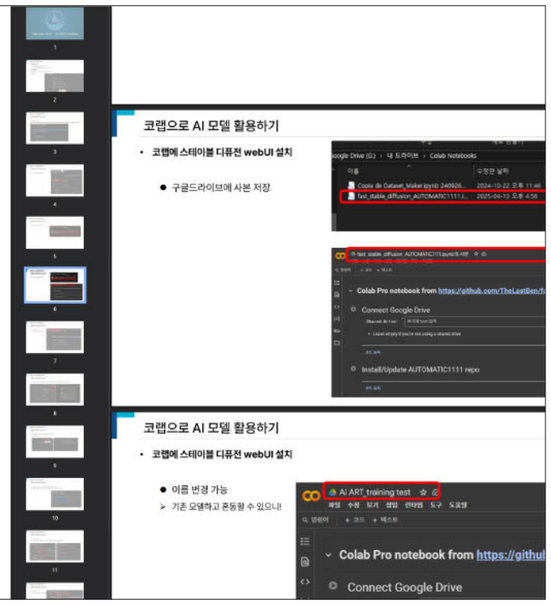
40대 남성, 약간의 수염, 긴 머리, 백인, 마른 편, 얼굴 정면 샷,
목판화

그런데...

캐릭터 일관성 유지 ChatGPT등에서 전혀 안되었는데...
개강하고 3주 뒤... ChatGPT...

일관성 유지 가능

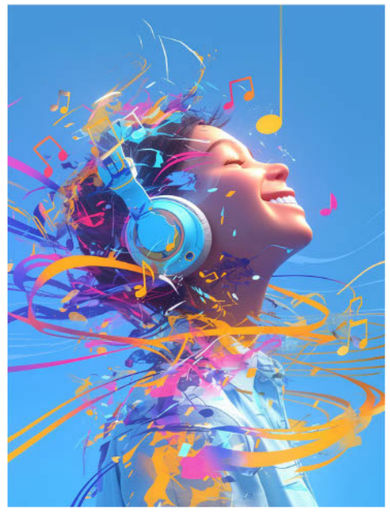
Web UI, Comfy UI의
의미가...



3. 학생 결과물 보고

이후 본격적으로
프로젝트 위주 수업으로 진행

각 학생의 결과물을
매주 발표, 비평, 수정



자신만의 음악 앨범 만들기

인간은 기획자
작사가는 인공지능
작곡가도 인공지능

중간프로젝트 결과물

싱귤래러티(Singularity) 그 자체

나보다 극도로 빠르게 진화하는 인공지능은
나보다는 빠르게 진화하는 학생들이 더 잘할지도...

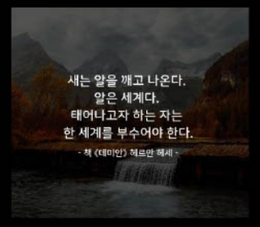
My Life is an Egg

김수연

앨범 기획

앨범명: Life is an Egg
https://youtu.be/v9XbfzQ4Sxs

앨범 전반 컨셉:
'삶'을 하나의 '알'(egg)로 비유해, 성장 과정 속에서 스스로 세계를 깨고 나오는 이야기를 다룬다. '삶은 계란'이라는 언어 유희를 통해, 유머와 따뜻함을 잃지 않으면서도 깊은 성찰을 담는다. 헤르만 헤세의 『데미안』 속 '알을 깨야 진정한 자아를 찾을 수 있다'는 메시지를 중심축으로 삼는다.



주제:
- 자신을 둘러싼 한계와 두려움을 깨고 성장하는 이야기
- 좌절, 상실, 사랑, 이별, 수용, 희망을 거쳐 가는 여정
- 결국 자신이 다시 스스로를 믿게 되는 순간까지

참고자료:
- 헤르만 헤세 『데미안』 (특히 "알을 깨고 나와야 한다"는 상징)
- 문보영 시집 『배틀그라운드』
- 개인적인 성장 서사, 모성애, 역설적 희망, 어린 시절의 순수함

Intro: My Paper Wings

곡 컨셉:
주인공의 태아기와 어린 시절을 꿈결처럼 묘사한다. 어머니의 뱃속에서 듣던 따뜻한 말들과 함께, 모든 것이 가능하다고 믿던 순수한 시기를 몽환적으로 표현한다. 마치 오래된 필름 속 유년기의 한 장면을 떠올리게 하는 사운드.

주제:
- 시작, 가능성, 무한한 꿈
- 아직 세계를 깨기 전의 무구함
- 따뜻한 사랑과 보호 속의 '알' 상태

사운드 이미지:
- 빛바랜 필름, 잔디밭, 비눗방울, 어린 아이의 웃음소리, 햇살
- 몽환적이고 부드러운 사운드 + 가벼운 여쿠스틱 악기들
- 중간중간 들려오는 어머니의 음성 (예: "건강하기만 하렴", "우리 아가야")

가사 방향성:
- 반복되는 어머니의 다정한 속삭임 (코러스처럼)
- "나는 선장이 될 거야", "세상에 행복함만 가득하게 할 거야" 같은 천진난만한 다짐들
- 모든 것이 가능할 것만 같은 낙관과 사랑

앨범 소개글 — *life is an egg*

삶은 하나의 알처럼, 여리고 투명한 껍질을 깨며 시작된다.
때로는 부서지고, 길을 잃고, 다시 일어선다.
사람을 보내고, 혼자서 꿈을 추고, 잊고 있던 정원으로 돌아간다.

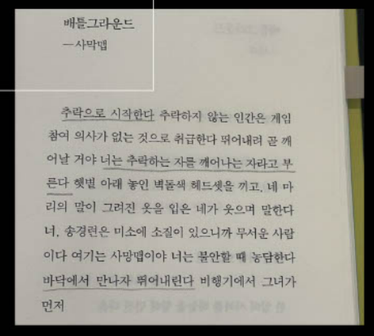
*life is an egg*는
그런 삶의 순간들을 종이로 접은 날개처럼 조심스럽게 담아낸 여정이다.
웃음과 눈물, 청묵과 고백이 뒤섞인 여섯 곡은
우리 모두의 작은 꿈과 작별, 그리고 다시 깨어남을 이야기한다.

"다시 깨어나는 법을 잊지 않기를."

My life is an egg

1. **intro — my paper wings**
   : 처음 세상을 만나던 날, 종이 날개를 달고 꿈을 꾸던 아이.
2. **my broken map**
   : 찢긴 지도 위를 헤매던, 길을 잃은 나의 기억.
3. **my worn-out wings**
   : 낡고 닳은 날개로도 계속 걸어야 했던 무채색의 하루.
4. **my all**
   : 사랑하는 사람에게 보내는 조용한 안녕, 슬픔을 숨긴 마지막 자장가.
5. **my midnight waltz**
   : 어설프고 웃긴 몸짓으로, 외로움을 춤추던 한밤의 왈츠.
6. **outro — back to my garden**
   : 문득 되살아난 오래된 정원, 그리고 다시 시작하는 작은 꿈.

My Broken Map

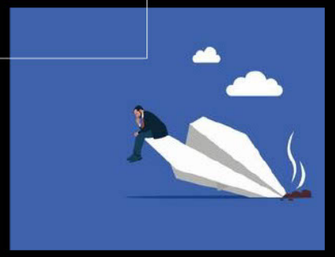


주제
- "추락으로 시작하는 성장통"
- 꿈이 무너지고, 사랑이 끝나고, 자신이 평범한 존재임을 처음으로 깨닫는 시기
- 외로운 마음, 그리워할 대상마저 사라진 공허함
- 그러나 감정은 담백하게 — 오히려 감정에 휩쓸리기보다는 "받아들이는" 무드

참고자료
- 문보영 시인 「배틀그라운드」: "추락으로 시작한다"는 구절
- 이별, 실패, 외로움을 다룰 때 터무니없이 드라마틱하기보다 조용히 무너지는 느낌을 살릴 것

사운드 이미지
- 느릿한 템포
- 통기타/피아노 베이스, 약간 허스키한 보컬 톤
- 반복되는 짧은 가사로 잔잔하게 몰입하게 만드는 스타일

My Worn – out Wings



곡 컨셉:

- 꿈이 무너진 후, 반복되는 일상 속을 떠도는 주인공.
- 특별한 기쁨도, 특별한 슬픔도 없이 '그냥' 살아가는 지친 하루들.
- 하지만 여전히 걷고 있고, 삶은 멈추지 않는다는 점을 포착.

노래 장르:

- Indie Jazz (또는 Soft Jazz) + Lo-fi 감성
- (반복적인 리듬, 부드럽고 무심한 톤)

노래 무드:

- Weary (지친),
- Resigned (체념한 듯한),
- Calm (잔잔한)

사운드 이미지:

- 빈 골목길에 혼자 걷는 듯한, 느긋하고 쓸쓸한 분위기.
- 작은 카페 스피커에서 흐르는 재즈처럼 자연스러운 무드.

My All

주제

- 어머니를 떠나보내기 직전, 주인공이 어머니 곁을 지키며 무서워하지 말라고, 괜찮다고 이야기하는 곡. 친구도 연인도 아닌, 시작부터 늘 곁에 있어주었던 어머니에게 보내는 마지막 인사. 어머니께 배운 언어로, 어린 시절 이야기와 농담을 섞으며, 자장가처럼 따뜻하고 조용한 분위기로 감싼다. 이별을 직시하기보다는, 평온하고 편안하게 잠들 수 있도록 곁을 지킨다.

장르

- soft lullaby folk / acoustic ballad

무드

- 따뜻함 / 잔잔한 슬픔 / 어린 시절의 향수 / 농담과 속삭임 / 부드러운 이별 / 조용한 용기

키워드

- 시작부터 곁에 있었던 존재, 어린 시절 놀이, 배운 언어, 자장가, 웃음 속 이별, 걱정 없는 잠

My Midnight Waltz

앨범 컨셉

- 잊을 수 없는 여름밤, 웃기지만 슬픈, 어설프지만 소중했던 한순간.
- 약간은 허무하고 해학적인 '밤의 춤'을 통한 기억과 감정의 기록.

곡 컨셉

- 한밤중 공원, 가로등 불빛 아래서 혼자 춤추는 모습.
- 외롭고 슬프지만, 동시에 웃긴 모습.
- 그 모든 게 '진심'이었기에 소중했던 기억.
- **"비루하지만 찬란했던 나만의 왈츠."**

노래 장르

- Indie Pop / Soft Jazz
- (경쾌한 리듬을 살짝 얹어주되, 멜랑콜리한 감성 유지)

노래 무드

- Whimsical (엉뚱한)
- Melancholic (쓸쓸한)
- Nostalgic (그리운)
- Playful (장난기 있는)

사운드 이미지

- 한밤중 살짝 흔들리는 가로등 불빛,
- 텅 빈 공원의 잔잔한 바람 소리,
- 귓가를 간질이는 옛 노래 한 조각.

Outro : Back to My Garden

앨범 컨셉

- 첫 곡 *My Paper Wings*와 연결되는 구조.
- 문득 스르르 잠에 빠진 듯, 과거의 따뜻한 기억을 마주한다.
- '나만 아는 조용한 비밀 정원'을 통해 희망을 다시 발견한다.

곡 컨셉

- 부드러운 드림팝/앰비언트 사운드.
- 영화 '마담 프루스트의 비밀 정원'
- 짧은 가사 + 풍성한 반주 중심. (반주가 서사를 끌어감)

노래 장르

- Dream Pop / Ambient Pop

노래 무드

- Dreamy (몽환적)
- Nostalgic (향수를 자극하는)
- Hopeful (희망적인)
- Gentle (부드럽고 다정한)

사운드 이미지

- 서서히 번지는 빛
- 나뭇잎 사이로 부는 바람
- 어릴 때 뛰놀던 정원의 한 조각

March Baby BLUES

이연우

https://www.youtube.com/watch?v=qTStdEihw7Q&ab_channel=%EC%9D%B4%EC%97%B0%EC%9A%B0

학생 대표작

RECOVER PROJECT

기말 프로젝트

- 자유 프로젝트
- 기본적으로 자신의 전공과 어울리는 작품 제시

프로젝트 과정

STEP 1

시집에 수록된 시 혹은 표제작을 바탕으로 분위기 키워드 추출

STEP 2

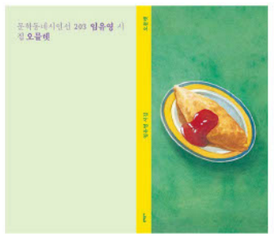
키워드를 기반으로 한 일러스트 생성

이때, 편집기준 설정을 통해 문집의 통일성을 살릴 수 있게 기획

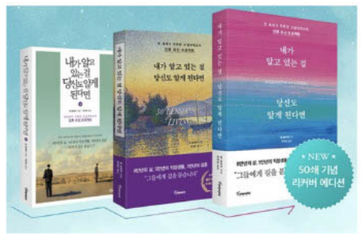
STEP 3

생성된 이미지를 표지로 활용

리커버 책의 유형



표지의 단어를 그대로 형상화



필수적 정보를 제외하곤 정보를 줄이고
일러스트의 강조

화려한 장식성과 강렬한 색상으로 꾸밈

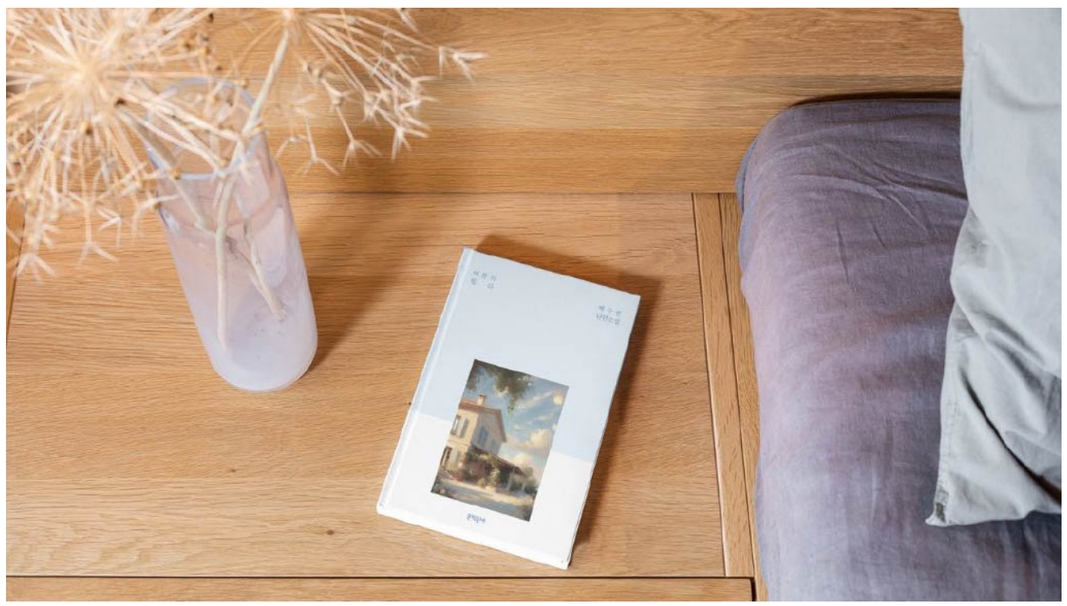
가장 용이한 1의 용례를 집중하여 프로젝트 진행

- 제목의 구운 두부를 활용하여 표지 제작.
- 생성형 AI를 활용하여 이전 레퍼런스(런치타임 리커버)와 기존 표지를 학습시킴
- 그 후 유사한 느낌으로 구운 두부를 그리게 하여 그를 표지로 활용
  이 외 요소는 기존 리커버 북의 형식을 그대로 따름

프로젝트 결과 1.
한여진, <두부를 구우면 겨울이 온다>

- 제목의 '여름의 빌라'를 활용하여 표지 제작.
- 생성형 AI를 활용하여 기존 표지와 유사한 그림체로
  "여름의 빌라"를 그리게 함
  (Chat GPT 활용 : 이 표지의 그림체와 색조를 유지하면서
  여름의 빌라를 그려줘)
- 이 외 요소는 생성형 AI의 추천을 바탕으로 이루어짐.

프로젝트 결과 2.
백수린, <여름의 빌라>



결론

"교수가 덜 건드릴 수록 훌륭하게 자라는 경우도 있다."

"올드스쿨의 미덕 그 자체의 의미는 존재"

왜냐하면 결국 생산된 콘텐츠의 가치는 결국
비효율적이고, 예측하기 힘들고 비정형적인 인간이라는 소비자를 통해 이루어지므로

"이 교과의 문제점?"

교과명, 교과코드가 같지만 내년엔 다른 교과

분과회의 세션 11
Parallel Session 11

AI와 한국어 교육
AI and Korean Language Education

한국어교육과 인공지능: 교사의 새로운 역할과 가능성

Korean Language Education and AI: New Roles and Possibilities for Teachers

이정희
경희대학교 교수

Junghee Lee
Professor, Kyung Hee University

초록

본 연구는 인공지능(AI)이 한국어교육에 미치는 영향과 이에 따른 교사 역할의 재정립을 탐색한다. 2005년부터 2025년까지 발표된 국내 학술 자료 148편(학술 논문 102편, 학위 논문 46편)을 대상으로 PRISMA 체계적 문헌 고찰과 LDA 주제모델링 기법을 적용하여 한국어교육 분야의 인공지능 관련 연구 동향을 분석하였다. 분석 결과, AI 기술 발전에 따라 '자동 채점 기반 쓰기 평가', '디지털 역량 기반 교수·학습 설계', 'AI 활용 쓰기 효과 검증', '대화형 AI 기반 말하기 연습'의 네 가지 주요 토픽이 도출되었으며, 이에 상응하는 교사의 역할을 메타 평가자로서의 교수자, 수업 설계자로서의 교수자, 피드백 코치로서의 교수자, 학습 조정자로서의 교수자로 제시하였다. 향후 한국어교육 분야에서는 한국어 교사의 디지털 리터러시와 AI 협업 역량 구인에 대한 구성 타당도 검증과 함께 이를 기반으로 한 측정도구 개발 및 구체적 내용 체계 구축이 필요할 것이다.

Abstract

The purpose of this study is to explore changes brought by artificial intelligence (AI) in the field of Korean language education and to investigate the new roles of teachers. Using PRISMA systematic literature review and LDA topic modeling, the research analyzed 148 domestic academic sources published between 2005 and 2025 (102 journal articles, 46 thesis and dissertations) related to AI in Korean language education. As a result, four topics were identified according to recent advances in AI technology: automated scoring-based writing assessment, digital competency-based instructional design, effect verification of AI-based writing practice, and conversational AI-based speaking exercises. Based on these findings, the study suggests evolving teacher roles as meta-evaluator, instructional designer, feedback coach, and learning coordinator. For this, it is necessary to verify the construct validity of Korean language teachers' digital literacy and AI collaboration competences in the field of Korean language education, to develop appropriate measurement tools based on this, and to create specific curricular content accordingly.

1. 서론

2024년 1월 미국 라스베이거스에서 개최된 CES(Consumer Electronics Show)에서 NVIDIA의 젠슨 황 대표는 인공지능 시대의 도래와 "물리적(Physical) AI"의 시대로 진입을 강조하면서 정보처리·추론·계획·행동이 통합된 인공지능 로봇에 대해 발표하였다. 교육의 특정 부분은 인간만이 유일하게 담당할 것이라고 생각했던 우리들에게 큰 도전이 시작된 것이다. 즉, 인공지능 교사의 역할이 지식 교육의 경계를 넘어설 수 있을지도 모른다는 불안과 두려움이 본격화되었다고 할 수 있다.

광범위하게 언어 교수에 적용되어 온 컴퓨터 지원 언어 학습(CALL: Computer-assisted language learning)은 컴퓨터를 언어 교수 및 학습에 적용하는 탐색 및 연구로 정의하고 있다(Levy, 1997: p. 1). CALL은 정보통신기술(ICT)을 활용한 다양한 교수·학습 방식과 접근을 아우르며, 1960~70년대의 전통적인 반복 연습(drill-and-practice)프로그램부터 시작하여 최근 가상 학습 환경(virtual learning environment), 모바일 보조 언어 학습(MALL) 등까지 광범위한 활용과 접근을 포함한다.

제2언어 습득에 있어 컴퓨터 지원 언어 학습이 성공하려면 외국어 교사들이 컴퓨터 작동법을 충분히 숙지하여 컴퓨터 자료 제작을 지도할 수 있어야 한다는 연구가 발표되었는데(Marty, 1982) 최근까지도 교사들이 컴퓨터 혹은 인공지능을 잘 이해하고 다룰 줄 알아야 한다는 논의들이 이어지고 있다. 이는 교수·학습의 주도권을 교사가 가지고 있다는 대전제에서 교수·학습의 효과를 증진시키기 위한 방안이나 도구로 컴퓨터나 인공지능을 활용한다고 생각했기 때문이다. 그러나 이제는 교실 내에서 교사가 주도권을 가지기 위해서 과연 기술을 따라가는 것이 우선적인가에 대해 반문해 보아야 하는 시점이라고 생각한다.

인공지능의 등장과 함께 언어교육을 접합한 연구들은 인공지능 보조 언어 학습(AI assisted language learning)로 명명하여 CALL의 하위 범주로 인식하고 있으나 보조, 조력의 개념을 넘어서는 인공지능과 인간 교사의 협력적인 관계를 구현해 내는 인공지능 협력 언어 교수 모형이 필요하다고 본다. 즉, AI-협력 언어 교육(AI-Collaborative Language Education, AICaLE)은 교사, 학습자 그리고 인공지능이 상호보완적으로 협력하여 교수·학습·평가·피드백의 전 과정을 설계하고 실행하고 성찰하는 언어교육의 패러다임으로 인공지능의 활용을 넘어 교사의 전문적이고 윤리적 판단과 공정성을 포함하는 개념이다.

이를 위해 한국어교육 분야에서 인공지능 기술에 대한 연구는 어떻게 이루어지고 있는지를 살펴보는 것은 매우 중요하다고 생각한다. 이를 통해 우리가 앞으로 정립해야 할 인간 교사와 인공지능의 역할이 무엇인지 논의해 볼 수 있을 것이다.

2. 연구 방법

2.1. 연구 대상

본 연구는 2005년부터 2025년까지 발표된 최근 20년간 한국어교육 분야의 인공지능 관련 논문을 수집하여 토픽 모델링(Topic Modeling)을 통한 텍스트 마이닝으로 전반적인 연구 동향을 분석하였다. 토픽 모델링은 다양한 문서 내 텍스트를 분석하여 내재된 주제를 파악하는 기법으로 본 연구에서는 토픽 모델링 알고리즘 중 널리 활용되는 LDA(Latent Dirichlet Allocation) 기법을 사용하였다. LDA는 문서가 임의의 수의 토픽으로 구성되어 있다는 가정하에 단어의 집합으로부터 토픽을 추출한다. 이는 디리클레 분포(Dirichlet Distribution)라는 연속 확률 분포를 활용하여 문서에서 명확하게 드러나지 않은 잠재적 주제를 통계적 방

법론으로 찾아내는 기법이다.

자료의 수집과 처리 과정은 PRISMA(https://www.prisma-statement.org/) 지침의 체계적 문헌 고찰(Systematic reviews) 과정을 따랐다. PRISMA(Preferred Reporting Items for Systematic reviews and Meta-Analyses) 지침은 체계적 문헌 고찰과 메타 분석을 위한 지침으로 독자를 위해 연구의 절차 이해를 돕고 연구의 투명성을 확보하기 위한 가이드 라인으로 제시되어 있다. 본 연구에서는 공개된 2020 버전의 체크리스트와 흐름도에 따라 연구를 진행하였다.
국내 학술연구정보 서비스인 RISS(riss.kr)에서 논문을 수집하였다. 2025년 9월 1일 기준으로 '인공지능', 혹은 'AI'와 '한국어교육'을 검색어로 설정하여 이 두 키워드를 모두 포함한 논문을 수집하였다. 학술 논문과 학위 논문을 포함하여 1차 수집은 총 443편이었으며 중복되거나 국어학, 국어교육, 외국어학, 언어학 등 한국어교육과 관련이 없는 논문을 제외하고 236편을 분석 대상으로 삼았다. 이후 원문과 서지 정보 등을 살펴 정성적으로 내용을 파악한 후 초록이 없거나, 주제와 맞지 않은 논문 51편을 제외하였다. 이후 원문을 확인하면서 내용이 한국어교육 분야가 아닌 연구 35편을 제외하여 학술 논문 102편, 학위 논문 46편 총 148편을 도출하였다.

2.2. 연구 절차

수집된 논문의 제목과 초록을 분석 기준으로 하여 전체 논문의 서지 정보를 추출하여 리스트로 만든 후 이를 토대로 텍스트 전처리 과정을 수행하였다. 파이썬 kiwipiepy 패키지를 이용하여 형태소를 분석한 후 일반적인 연구 용어와 본 연구의 주제어를 불용어 처리하였다.[1] 이후 주제와 관련한 내용어만 선별하기 위해 형태소 태그 nn*(NNG, NNP)이 붙은 단어만을 추출하였으며 의존 명사 등을 포함하지 않기 위해 1음절 단어는 제외하였다.

전처리를 마친 데이터는 파이썬의 gensim 라이브러리를 통해 LDA 알고리즘을 거쳐 주제를 추출하였다. 사용자가 토픽 수를 지정해야 하므로 토픽 수별로 응집도(coherence)와 복잡도(perplexity)의 지수를 고려하였으며 전체 데이터 양과 응집도의 엘보우 포인트(응집도가 증가에서 감소하는 지점)를 고려하여 최종 토픽 수를 4개로 설정하였다.[2]

LDA 기법은 출현 확률에 따라 설정된 토픽 수에 따라 키워드를 분류하는 기법이므로 결과를 한눈에 알기 쉽도록 군집과 키워드를 함께 나타내는 LDAvis 시각화를 사용하였다. 또한 4개의 토픽은 생성형 인공지능을 활용하여 키워드 목록을 통해 라벨링하였다. 그리고 연도별 연구에서의 토픽을 시계열 그래프로 나타내어 연구의 흐름을 파악하였다. 전체 연구 절차는 자료 수집, 자료 선별, 전처리, 토픽 추출, 결과 분석의 5단계로 구성된다.

1) 연구, 분석, 방안, 중심, 관련, 필요, 가능, 기반, 제시, 방법, 결과, 대상, 논문, 분야, 목적, 중요, 조사, 제안, 내용, 인식, 바탕, 방식, 작업, 본고, 논의, 수행, 때문, 방향, 결론, 한국어교육, 인공지능, 인공, 지능, ai, 한국어, 교육

2) 응집도(coherence)와 복잡도(perplexity)는 토픽 수 3개, 4개, 5개일 때 각각 (0.29. -6.08), (0.34, -6.03), (0.35, -6.02)로 나타났다. 응집도가 높을수록, 복잡도가 낮을수록 일반적으로 적정 토픽 수라고 볼 수 있어 토픽을 5개로 선정하는 게 가장 적합하다고 볼 수 있으나 토픽 3과 토픽 4 사이에 응집도가 큰 폭으로 상승하여 토픽 주제에 대한 개선이 크게 향상한 것과 토픽 4와 토픽 5의 수치가 크지 않은 점도 고려하여 토픽 4개로 결정하였다. 그리고 본 연구는 동향 분석을 바탕으로 한 교사 관련 논의를 다루어 교사 관련 키워드를 추가 분석해야 하는데 토픽 5개로 구성할 경우 교사 키워드 분석에 고유 키워드가 일부 토픽에서 나타나지 않는 한계가 도출되어 적정 교사 관련 토픽 수를 확보하기 위해 토픽 4개로 선택하여 분석하였다.

3. 연구 결과

3.1. 연도별 동향 분석

분석 대상의 논문은 시기를 설정하지 않고 결과를 검색하였다. 한국어교육에서 최초의 인공지능 관련 연구는 조수진(2005)에서 시작한 것으로 확인되었다. 이후 약 15년간 관련 연구가 거의 이루어지지 않다가 2016년 1월 스위스 다보스 세계경제포럼(WEF)에서 제4차 산업혁명이라는 용어가 주목받은 후 국내에서는 2019년 3편, 2020년 2편의 연구가 발표되었는데 이 중 3편이 가상현실(VR)과 관련한 논의이다. 2019년 코로나 팬데믹 이후 2021년과 2022년에 각각 10편과 11편으로 연구물들이 증가하여 이 시기에 한국어교육에서의 인공지능에 대한 학술적 관심이 높아진 것으로 나타났다.

2022년 11월 chatGPT 3.5가 본격적으로 서비스를 제공하면서 2023년에는 전년의 두 배가 넘는 연구인 23편이 발표되었으며 지금까지 증가세를 이어 가고 있다. 논문 유형으로는 전체의 약 69% 이상을 학술 논문이 차지하여 새로운 이론의 개발이나 모델 제안과 관련한 연구보다는 소규모의 사례 분석 등의 실증적인 연구가 논문의 중심이 되는 것을 확인할 수 있었다. 특히 2025년에 들어 박사 학위 논문이 7편으로 급증하여 인공지능 관련 관심이 증대되었다는 것을 알 수 있다. 연구 주제 역시 한국어 쓰기 교육과 관련한 연구가 3편으로 많지만 말하기 교육, 어휘 교육, 교사 교육, 메타버스 설계와 같이 다양한 분야로 연구의 범위가 확대되었다는 것을 확인할 수 있다. 이러한 박사 학위 논문의 증가는 앞으로 AI 분야의 전문가로 확장될 개연성이 높아 연구의 저변 확대로 인해 다양한 연구가 증가할 것으로 보인다.

3.2. 토픽 모델링 분석

토픽 모델링 분석은 응집도(coherence)와 복잡도(perplexity)를 고려하여 크게 네 개의 군집으로 나누어 살폈다. 분석 결과는 시각화 기법 중 하나인 LDAvis를 활용하여 도식화하였으며 각 토픽에 나온 상위 키워드를 바탕으로 생성형 인공지능 서비스를 통해 주제명을 라벨링하였다.[3] 그 결과 토픽 1은 '자동 채점 기반 쓰기 평가', 토픽 2는 '디지털 역량 기반 교수·학습 설계', 토픽 3은 'AI 활용 쓰기 효과 검증', 토픽 4는 '대화형 AI 기반 말하기 연습'으로 명명하였다.

	토픽 1	토픽 2	토픽 3	토픽 4
주제명	자동 채점 기반 쓰기 평가	디지털 역량 기반 교수·학습 설계	AI 활용 쓰기 효과 검증	대화형 AI 기반 말하기 연습
비중	19.8%	31.5%	21.9%	26.9%

표1 토픽 4개에 대한 토픽 모델링 분석 결과

전체 분석 논문의 연도별 추이를 살펴보면 다음과 같다. 2005년 1편의 연구가 발표된 후 2019년 이전까지 있던 연구 공백은 추세 변화에서 제외하여 연도 분석은 2019년부터 2025년까지 실시하였다. 각 연도별 토픽 4개의 비중을 살펴보면 다음 그림 3과 같다.

3) 대화형 인공지능 서비스인 클로드(claude) Opus 4.1, chatGPT5의 도움을 받아 저자가 적절한 단어로 수정하였다.

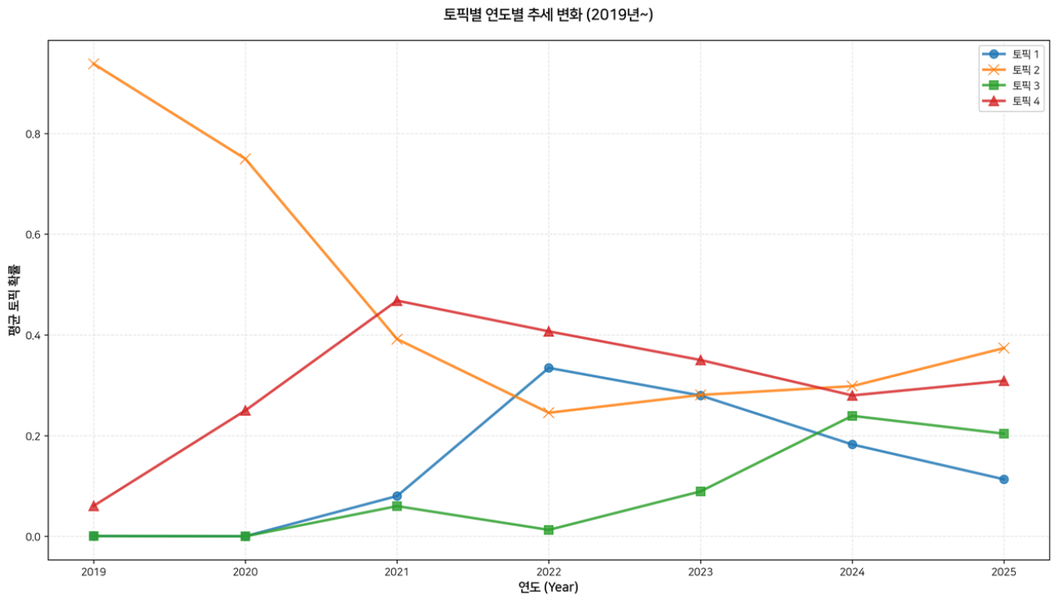


그림2 2019년 이후 토픽별 연도별 추세 변화

자동 채점 모델의 토픽 1은 눈에 띄는 증감 흐름을 보이지는 않았다. 한국어교육에서 2020년 이후 주제에 대해 주목하기 시작한 것으로 보이나 최근까지는 인공지능 관련으로 한국어교육에서의 논의가 활발하지 않다. 디지털 역량 기반 교수·학습 설계의 토픽 2는 2019년에 인공지능 관련 가장 높은 관심을 받았으나 점차 유의미한 감소 추세를 보이고 있다. AI 활용 쓰기 효과 검증인 토픽 3은 4개의 토픽 중 유일하게 유의미한 증가세를 보이고 있다.[4] 2023년에서 2024년 사이에 연구물이 가장 큰 폭으로 증가한 것은 chatGPT 등의 대화형 인공지능 서비스를 활용한 연구로 집중된 결과로 보인다. 대화형 AI 기반 말하기 연습의 토픽 4는 2021년까지 증가하다가 하락하고 있는 것을 확인할 수 있다. 이는 팬데믹으로 인해 비대면 학습의 필요성이 대두되면서 디지털 활용의 챗봇 개발 논의가 시작되었으나 대화형 인공지능 서비스가 대중화되면서 연구 주제가 이동한 것으로 보인다.

토픽 추이 분석 결과, 일부 토픽에서는 뚜렷한 증감 추세를 보이지는 않았지만 뚜렷한 주제 전환 기점을 확인할 수 있었다. 디지털 교수 설계(토픽 2)는 2019년을 정점으로 감소하였고, 대화형 학습(토픽 4)은 코로나19 시기인 2021년 이후 정체되었다. 반면 쓰기 피드백 분석(토픽 3)은 생성형 인공지능의 대중화와 함께 2023년 이후 유의미하게 증가하고 있으며 현재 가장 활발한 연구 영역으로 발전하였다.

4) 토픽 4개에 대한 증감에 대한 회귀 분석에서 p값이 각각 토픽 1(0.218), 토픽 2(0.038), 토픽 3(0.009), 토픽 4(0.366)로 나타나 ($p<0.05$) 기준에서 토픽 2의 감소와 토픽 3의 증가만 유의미한 결과라고 볼 수 있다. 토픽 1과 토픽 4는 유의미한 증감을 보이지 않아 뚜렷한 경향성이 없다고 판단하였다.

3.3. 토픽별 세부 키워드 분석

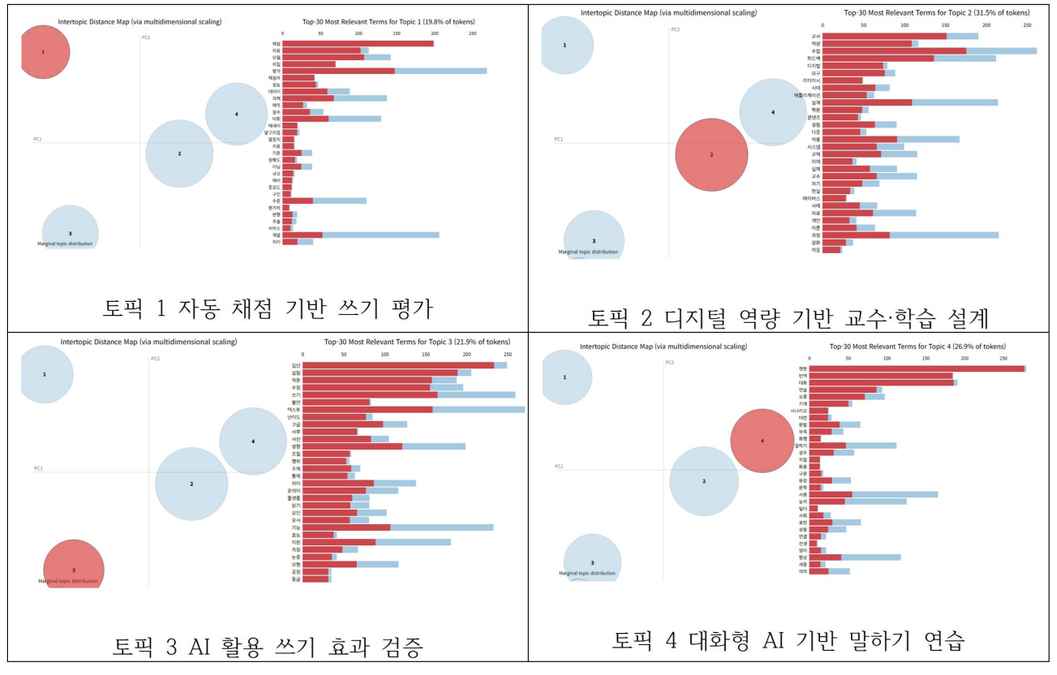


그림 2 토픽별 LDAvis 시각화 결과(λ=0.4)

(1) 토픽 1: 자동 채점 기반 쓰기 평가

토픽 1은 자동 채점 기반 쓰기 평가의 연구 내용을 담고 있으며 '채점', '평가', '모델', '자동', '자질', '과제', '어휘', '데이터', '개발', '사용', '성능', '과정'의 순으로 핵심 키워드 출현하였다. 전체 토픽에서 19.8%의 가장 적은 비중을 차지하였으며 이 토픽은 쓰기 자동 채점, 읽기 텍스트 난이도 분석 등 주로 읽기와 쓰기 등 문어와 관련되어 자동 평가 및 난이도 예측에 관한 자동 시스템 및 검증과 관련된 주제이다. 또한 앞으로의 자동 채점 모델이 공정성과 타당성을 확보하기 위해 어떤 자질이 필요한지에 대한 연구들로 앞으로의 교육적 평가 시스템 설계에 매우 중요한 의의를 갖는 주제이다. LDAvis 내 군집 형성 위치를 보면 2, 3, 4의 토픽과 뚜렷하게 구분되어 완전히 독립적인 위치를 잡고 있다. 이것은 가장 적은 비중을 차지하는 것과 맞물려 아직까지 교수·학습의 관점에서 실제적인 적용으로의 연구가 진행되지 못하고 검증 단계에서 머물고 있다는 한계가 드러난 것이라고 볼 수 있다.

(2) 토픽 2: 디지털 역량 기반 교수·학습 설계

토픽 2는 디지털 역량 기반 교수·학습 설계 주제로 '수업', '교사', '피드백', '설계', '역량', '적용', '개발', '과정', '요구', '디지털', '교재', '교수'의 순으로 키워드가 나타났으며 전체 토픽 중 가장 큰 비중(31.5%)를 차지하고 있다. 조수진(2005)를 비롯해 초기 연구들이 토픽 2와의 연관이 높은 것으로 나타났다. 개별 애플리케이션나 웹 기반 교재 시스템, 매체 활용 등 학습 목적에 따라 교수·학습 설계를 위해 인공지능 플랫폼을 활용하는 연구가 중심이 되었다. 이를 위해 실제 도구로 어떻게 활용해야 하는지에 대해 입출력을 중심으로 한 구성에 주목하였다. 그 가운데 생성형 인공지능이 대중화된 2023년 이후에는 실제 애플리케이션 및 플랫폼 개발

의 기초 연구에서 인공지능 활용에 초점을 두어 사용 경험 및 개발 사례의 효과 분석 등으로 학습 도구 응용과 사례 연구를 통한 교수 설계 측면으로 연구 방향이 전환되었다.

(3) 토픽 3: AI 활용 쓰기 효과 검증

토픽 3은 AI 활용 쓰기 효과 검증으로 '집단', '실험', '쓰기', '텍스트', '작문', '수정', '평가', '영향', '기능', '사용' 순으로 키워드가 나타났으며 전체 토픽 중 21.9%의 비중을 차지하였다. 토픽 3은 쓰기라는 특정 언어 기능에 집중되어 인공지능 기술이 쓰기 교육의 보조 역할로 어떻게 활용되는지에 대한 주제이다. 학습 도구를 사용할 때의 학습자의 쓰기 효능감과 불안, 자기주도 학습의 영향 등 실제 피드백을 통한 효과에 저해하는 요인 등이나 인공지능 기술의 학습 도구에 대한 인식 정도에 대한 연구 등으로 실제 인공지능을 통한 언어 중심의 교육적 효과 검증에 대한 연구는 상대적으로 미비하다. 이와 같은 결과로 LDAvis에서의 토픽 3의 군집 위치도 실제 교수·학습과 관련된 토픽 2(디지털 역량 기반 교수·학습 설계)와 토픽 4(대화형 AI 기반 말하기 연습)과도 상대적으로 먼 거리에 있다. 토픽 1과 마찬가지로 학습 도구로의 인공지능 피드백 연구가 실제 학습의 효과 검증 및 작문 내용 분석에 대한 연구가 필요할 것이다.

(4) 토픽 4: 대화형 AI 기반 말하기 연습

토픽 4는 대화형 AI 기반 말하기 연습에 대한 주제로 '챗봇', '대화', '번역', '연습', '오류', '사용', '기계', '개발', '말하기', '능력' 순으로 키워드가 나타났다. 토픽 2와 토픽 4는 실제적으로 한국어교육에서 교수·학습에 활용되는 부분을 다루고 있어 LDAvis의 군집 위치가 상대적으로 가깝게 나타났다. 토픽 2가 교수 설계에 중점을 두어 그 실제적인 활용 방법을 모색하고 그 효과를 검증하는 주제를 나타낸다면 토픽 4는 챗봇 및 생성형 인공지능 서비스를 바탕으로 한 학습자와의 상호작용에서의 학습 효과에 중점을 두고 있다는 데 차별점이 있다. 따라서 '말하기', '화용', '상황', '사용', '표현' 등 실제 언어 사용에 주목하고 있다. 이는 토픽 3이 쓰기라는 언어 기능에 집중한 것과 비교하여 토픽4는 말하기를 포함하여 가장 폭넓게 인공지능 기술을 활용하는 연구 주제라는 것을 의미한다. 언어 기능이 특정 영역에 제한되지 않아 앞으로의 연구 가능성이 많다고 볼 수 있지만 앞서 살핀 연구 흐름에서는 뚜렷한 증가세를 보이지 않았다. 이는 코로나 19에서의 비대면에 대비한 시기적 관심과 맞물려 챗봇 개발 및 적용에 대한 연구가 증가했다가 생성형 인공지능 서비스를 활용한 대화형 학습이 대중화되면서 교육적 활용 도구가 바뀐 것이 원인으로 보인다. 최근에는 chatGPT 등의 대규모 언어 모델을 기반으로 한 대화형 학습을 통한 연구가 주를 이루고 있다.

3.4. 교사 관련 키워드 분석

토픽 모델링을 통해 도출한 4개의 토픽에 대해 토픽 내에서 다시 교사 관련 키워드를 뽑아 분석하였다. 토픽별로 분류된 주요 키워드 중 '교사', '교원', '강사', '지도', '강의', '안내', '연수', '훈련', '전문성', '코칭', '설계', '채점자', '전문가' 등과 같이 교사와 관련되었다고 볼 수 있는 키워드만을 선별하였다. 선별 키워드를 통해 각 토픽 내에서 토픽 고유의 교사 관련 키워드의 출현 확률을 바탕으로 하여 다음 표 2와 같이 토픽별 필요한 교사의 역할을 정의하였다.

	주제명	교사 역할(비)	교사 관련 주요 키워드
토픽 1	자동 채점 기반 쓰기 평가	메타 평가자(17.10%)	채점, 평가, 자질
토픽 2	디지털 역량 기반 교수·학습 설계	수업 설계자(30.80%)	수업, 교사, 역량
토픽 3	AI 활용 쓰기 효과 검증	피드백 코치(25.90%)	지원, 도움, 피드백
토픽 4	대화형 AI 기반 말하기 연습	학습 조정자(26.40%)	대화, 연습, 응답

표 2 토픽별 교사 역할 정의 및 주요 관련 키워드

(1) 토픽 1: 메타 평가자

토픽 1은 '채점', '평가', '자질'이 가장 높은 교사 관련 키워드로 나타났다. 이밖에 다른 토픽에 나타나지 않은 고유 키워드를 보면 '채점자'와 '검토'가 있다.[5] 따라서 교사는 학습자를 평가하기 위해 문항을 개발하고 이에 대해 채점 점수를 그대로 평가에 반영하는 전통적인 평가자의 역할에서 더 나아가 평가 자체에 대한 전문성을 요구한다. 자동 채점 모델을 통해 자동 문항과 개발과 결과를 도출하면 그 결과의 공정성과 타당성을 검토해야 하며 평가 기초의 기준과 루브릭을 비판적으로 점검할 수 있어야 한다. 따라서 평가에 대한 메타적 인식과 지식이 앞으로 더 요구될 것으로 보인다.

(2) 토픽 2: 수업 설계자

토픽 2에서 교사 관련 주요 키워드는 '수업', '교사', '역량'으로 꼽혔다. 이밖에 고유 키워드로 '콘텐츠'를 포함하여 실제 학습 목표에 맞게 인공지능 기술을 활용하여 콘텐츠를 제작하는 설계자의 역할을 강조하고 있다. 따라서 교사는 수업의 의도를 명확히 파악하고 이를 학습자에 맞게 조정하고 과제를 제시할 수 있어야 한다. 이는 기존의 수업 설계자로서의 교사 역할에서 벗어나지 않지만 인공지능 기술을 활용한다는 점에서 기술을 교육에 활용할 때의 장단점을 파악하고 이에 대한 활용 범위와 수업에서의 위치를 알아야 한다. 교사는 수업 설계자로서 수업에서의 인공지능의 역할을 결정하고 운영 지침을 제시할 수 있어야 한다(Pokrivčáková, 2019).

(3) 토픽 3: 피드백 코치

토픽 3은 쓰기 피드백 분석으로 '지원', '도움', '피드백'의 주요 키워드를 포함하였으며 고유 키워드가 없이 대부분의 교사 관련 키워드가 다른 토픽에서도 출현하였다. 인공지능 기술을 활용하면 학습자의 산출물에 대한 교정 작업은 자동으로 처리할 수 있다. 따라서 기존에 교사가 학습자의 산출물을 기반으로 한 피드백 역할을 하던 것을 기술이 담당하게 되는데 이때 교사는 자동 피드백된 내용에 대해 점검하고 분석하여 결과에 대한 타당도와 신뢰도 문제를 보완하는 역할을 해야 할 것이다. 즉, 인공지능 기술의 교정 결과가 기준에서 벗어나 과도하거나 미약한지를 판단하고 이것이 학습자의 정의적 측면을 함께 고려하여 올바른 효과를 나타낼 수 있도록 감시해야 한다.

5) 타 토픽에 나타날 확률이 0.00인 값을 나타내는 키워드이다.

(4) 토픽 4: 학습 조정자

토픽 4는 '대화', '연습', '응답'이 교사 관련 주요 키워드로 나타났다. 상위 키워드 3개가 모두 타 토픽에서 출현하지 않은 고유 키워드이다. 대화형 학습을 통한 상호작용을 중심으로 학습자의 언어 능력 향상이라는 명확한 주제가 있어 학습 내용에 맥락과 상황이 핵심 역할을 한다. 따라서 교사는 인공지능을 반영하지만 가장 현실에 가까운 연습과 과제를 구성할 수 있어야 한다. 또한 인공지능 학습 도구의 성능을 관리하고 이를 학습에 맞게 조정할 줄 알아야 한다. 그러기 위해서는 인간과 기술이 분리되는 기준이 무엇인지에 대한 윤리적 기준을 가지고 적절하게 판단할 수 있어야 한다.

4. 결론

이상의 분석에서 알 수 있듯이 교육의 다양한 맥락에서 인공지능은 떼려야 뗄 수 없는 중요한 요소라는 것을 확인할 수 있다. 특히 메타 평가자로서의 교수자, 수업 설계자로서의 교수자, 피드백 코치로서의 교수자, 학습 조정자로서의 교수자의 역할 변화에 따른 교수자 역할에 대한 깊이 있는 탐색과 논의가 이루어져야 할 것이다. 그러나 그간의 연구들은 학습자에 집중되어 인공지능의 학습 활용성과 학습 결과에 대한 효용에 대한 언급이 주를 이루며 실제 교실에서의 수업을 설계하고 실행하는 주체로서의 교사의 인공지능 역량 혹은 디지털 역량과 교사의 역할에 대한 논의는 미미하다.

연구의 방향성은 국제적이고 최신 기술 등의 변화와 함께 고찰되어야 한다. Gartner Hype Cycle은 2025년 주요 메가트렌드를 AI 에이전트의 부상, AI 레디 데이터와 데이터 관리 혁신, 멀티모달 AI의 필수화, AI 신뢰·위험·보안 관리와 윤리, 생성형 AI의 본격 산업 적용을 꼽고 있다. 기술적인 변화에 맞추어 고등교육의 경향을 발표한 Horizon Report(2025 EDUCAUSE)에 따르면 AI 기반 교수·학습 도구, 생성형 AI에 대한 교수 역량 개발, AI 거버넌스와 신뢰성, 윤리, 보안, 데이터 보호, 사이버 보안 강화, 진화하는 교수 방법, 디지털 리터러시 및 비판적 디지털 리터러시를 핵심 기술 및 실천으로 꼽고 있다.

한국어교육을 포함한 모든 교육 현장에서 관찰되는 핵심 기술이 인공지능이라는 것은 분명한 사실이다. 이에 인공지능 기술의 발전과 심화에 대해 윤리적 판단과 비판적 사고를 할 수 있는 교수·학습 주체로서의 교사의 역할과 교육에 대한 논의는 매우 시급하다. UNESCO에서 2024년에 출간한 "AI competency framework for teachers"에서는 인공지능 역량의 핵심 영역으로 인간 중심적 사고, 인공지능 윤리, 인공지능 기초 및 응용, 인공지능 교육학, 인공지능을 위한 전문성 개발로 구분하고 있다. 인공지능 역량의 가장 첫 번째를 인공지능 기술을 비판적으로 이해하고 인간 복지를 증진하는 데 사용할 수 있어야 한다는 대전제를 제시하고 두 번째로 인공지능 사용에 따른 윤리적, 사회문화적, 환경적 문제에 대해 깊이 있는 이해를 가지고 문제를 해결할 수 있는 능력을 제시하고 있다. 한국어교육 분야에서도 한국어 교사의 디지털 리터러시와 AI 협업 역량의 구성 타당도를 검증하고 이를 기반으로 한 측정 도구의 개발과 함께 구체적인 내용 개발이 이루어져야 할 것이다.

<참고문헌>

조수진(2005), 人工知能型 韓國語 말하기 코스웨어 開發 硏究, 어문연구 33-4, 한국어문교육연구회, pp. 527-546

Bekiaridis, G., & Attwell, G. (2024). Supplement to the DigCompEDU Framework: Outlining the skills and competences of educators related to AI in education. AI Pioneers Project. https://aipioneers.org/wp-content/uploads/2024/01/WP3_Supplement_to_the_DigCompEDU_English.pdf

Gartner, Inc. (2025). Gartner Hype Cycle for Emerging Technologies, 2025. Stamford, CT: Gartner, Inc.

Levy, M. (1997). Computer-assisted language learning: Context and conceptualization. Oxford: Clarendon Press.

Marty, F. (1982). Reflections on the use of computers in second language acquisition — II. System, 10(1), 1–11. 10.1016/0346-251X(81)90062-2

Pelletier, K., McCormack, M., Reeves, J., Robert, J., & Arbino, N. (2025). 2025 EDUCAUSE Horizon Report: Teaching and learning edition. EDUCAUSE.

Pokrivčáková, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. Journal of language and cultural education.

UNESCO. (2024). AI competency framework for teachers. United Nations Educational, Scientific and Cultural Organization.
https://unesdoc.unesco.org/ark:/48223/pf0000391104

인공지능 시대 인도를 위한 한국어: 기회와 도전

Korean for Indians in the Age of AI: Opportunities and Challenges

사티안슈 스리바스타바
네루대학교 교수

Srivastava Satyanshu
Professor, Jawaharlal Nehru University

Abstract

The growing popularity of Korean language in India reflects expanding cultural and academic exchanges between the two nations. However, India's multilingual environment, spanning Indo-Aryan, Dravidian, Tibeto-Burman, and English-speaking learners, poses complex pedagogical challenges. This paper examines how Artificial Intelligence (AI) can enhance Korean language teaching in India through adaptive, personalized, and interactive learning models. Drawing on research in Intelligent Computer-Assisted Language Learning (ICALL) and Natural Language Processing (NLP), it discusses how AI tools can address pronunciation, grammar, and translation challenges across diverse learner groups. At the same time, it highlights ethical, technical, and policy issues related to AI adoption, including the digital divide and cultural nuance loss. The paper proposes a contextualized framework for integrating AI responsibly in Korean language pedagogy, emphasizing teacher–AI collaboration and India–Korea academic cooperation.

Keywords: Korean language education, Artificial Intelligence, multilingual learners, India, pedagogy

Popularity of Korean Language in India:

The popularity of the Korean language in India has witnessed an unprecedented surge over the past two decades, reflecting the expanding scope of India–Korea relations in diplomacy, education, business, and culture. The India-Korea bilateral consular relations were established in 1962 and further strengthened with the appointment of an ambassador in 1973. Over the years, the bilateral ties have made significant progress and have become multidimensional, driven by shared interests, goodwill, and high-level exchanges. The opening of Indian economy in 1990s made way for Korean companies to start operating in India which in turn led to rise in the demand for Korean language as students saw an opportunity of employment.

Institutionally, Korean language education in India has evolved from a niche academic pursuit to a structured discipline. Korean language education in India began early in 1970s with a certificate and diploma course at Jawaharlal Nehru University (JNU) in New Delhi which was later upgraded to a degree Programme in 1995. JNU remains the flagship institution, with several universities and private academies now offering Korean as part of undergraduate and certificate-level programs. Currently, there are 5 central universities in India that offer degree Programmes and over 20 institutions offering certificate courses in Korean. India currently has a total of 6 functional King Sejong Institutes one each at the Korean Cultural Center in New Delhi (Delhi), Patna (Bihar), Imphal (Manipur), Barasat (West Bengal) and two centers at Chennai (Tamil Nadu).

Korean language also got included in India's New Education Policy (2020) as a new foreign language at secondary level which has further added to the popularity of the Korean language in India and is expected to further boost the demand of Korean language experts, positioning Korean as one of the most dynamic foreign languages in the Indian educational landscape.

India's Multilingual Classrooms:

The number of Indian learners of Korean has been growing rapidly in recent years. This is due to a number of factors, including the increasing popularity of Korean culture and media, the growing economic ties between India and South Korea, and the increasing number of scholarships and study abroad opportunities available to Indian students.

According to UNESCO's 2009 Linguistic Diversity Index, India is one of the world's most linguistically diverse nations. According to the 2001 census, there are 121 languages spoken by more than 10,000 people, and 22 of these are acknowledged as official languages under the constitution (Government of India 2001). This linguistic diversity is reflected in Indian classrooms across the country, where multiple languages are used by teachers and learners.

India's linguistic diversity profoundly shapes how Korean is learned and taught. The Indian classroom is inherently multilingual, encompassing learners from Indo-Aryan (e.g., Hindi, Bengali, Marathi), Dravidian (e.g., Tamil, Telugu, Kannada), and Tibeto-Burman (e.g., Manipuri, Bodo) language families, alongside English-medium students. Each group brings distinct phonological, morphological, and syntactic frameworks that influence Korean acquisition.

For example, Dravidian languages share typological similarities with Korean, such as agglutinative morphology and subject–object–verb (SOV) sentence structure, which can facilitate grammar learning. Conversely, Indo-Aryan speakers often rely on English as a mediating language, making the learning process translation-heavy and more dependent on contrastive analysis. Moreover, the sociolinguistic reality of India, where code-switching and translanguaging are common, creates both opportunities and complications in establishing Korean as a stable linguistic system within the learner's repertoire.

The constitution of Indian learners is also unique as they come from varied language and cultural backgrounds. For instance, a typical Korean language classroom in India may include students from five to ten different language and cultural backgrounds. Hence the mode of instruction in the classroom plays a crucial part in learning Korean.

Recognizing these heterogeneous learner backgrounds is essential for developing pedagogical strategies that move beyond a monolingual model. AI-driven adaptive systems, if properly trained, could potentially account for these variations and provide differentiated instruction tailored to the linguistic and cognitive profile of each learner.

Current State of Technology Usage in Korean Language Education in India

The integration of technology in Korean language education in India remains at a formative stage. While digital platforms such as Duolingo, TalkToMeInKorean, and Naver's Papago translation app are widely used by independent learners, institutional adoption of technology-assisted pedagogy is still limited. Technology usage is restricted to language lab (if applicable) or audio-visual system assistance and mainstream pedagogy still relies on the classical method or the grammar translation method. This means:

- The instruction in classes is conducted in English.
- Emphasis is placed on memorizing isolated vocabulary and grammar rules.
- Reading is the primary focus, accompanied by grammatical analysis.
- Translation exercises, usually from the second language L2 to English, are conducted.
- Little to no attention is given to oral production.

However, recent developments indicate a gradual shift. The COVID-19 pandemic accelerated the digital transformation of language education, prompting instructors to experiment with online learning management systems (LMS) like Google Classroom, Zoom, and Moodle. These tools, though not AI-based, have laid the foundation for integrating more sophisticated technologies. The use of ChatGPT and other generative AI tools is emerging informally among students and teachers for grammar practice, vocabulary explanation, and translation exercises, albeit without structured pedagogical frameworks.

At present, the main challenge lies in moving from technology-assisted learning to AI-integrated learning. The absence of localized AI models trained on Korean–Indian language pairs, inadequate digital infrastructure, and limited teacher training in AI pedagogy hinder large-scale implementation. However, it is being felt by the educators and learners alike that AI will prove to be a transformative force in the next phase of Korean Language education in India.

The Potential of AI Integration in the Indian Context

The advancements in technology have brought significant, widespread, and intricate changes to language learning, assessment, and research. Furthermore, these changes are expected to continue growing, as technology becomes increasingly integrated into the lives of language learners, both in formal education settings and their everyday lives.

The initial utilization of computer technology in foreign language teaching dates back to the 1960s, wherein mainframe computers were employed within a Skinnerian behaviourist approach. Learning a language at that time involved memorizing a set of predetermined responses, including frequently used vocabulary items, clichés, and phrases appropriate for specific conversations. Such approach is often referred as Computer-assisted language learning (CALL). According to Blake (2008), in the absence of a study abroad programme, technology has the potential to significantly improve L2 learners' interaction with the target language if utilised carefully. Traditional classes when combined with technology can become more interesting and can create a lot of excitement in language learners. Kern (2006) argues that the use of CALL is advantageous when it comes to teaching language in some context, hence its use in language education is quite obvious. He adds that the interactive nature of the computer creates a window for interaction and the learner benefits from the use of language based on the context of the interaction.

Since 2020, with the advent of technologies like 5G, new educational platforms are being created. Among them the most talked about is the Virtual Reality (VR) class taking place in a Metaverse. Especially, the 'Speaking class' using VR is gaining a lot of traction for its effectiveness and immersive experience whereby the learner can practice in a real-like

environment. This technology of course opens up great opportunities and possibilities for foreign language learners who are learning from their native countries. The representative example of a VR tool for Korean language practice is 'ImmerseMe'. It is a virtual reality-based language tool for students, travelers and business people. The learner enters a virtual reality zone and interacts with a native character to practice speaking or completing tasks. Shim and Yu (2019) call such VR-based language education a prime example of CALL (Computer-Aided Language Learning) and the core of which lies in the Natural Language Processing (NLP)158 technology. Further, many other Korean education services are being developed today that use NLP in Chatbots, AI Translation tools etc. With the fourth industrial revolution and technological advancements, it is definitely a matter to think about how we can efficiently apply such technologies to the Korean language education industry.

Opportunities & Challenges

Opportunities

The emergence of Artificial Intelligence (AI) in education has created unprecedented opportunities to reimagine foreign language pedagogy. For Korean language education in India, where learner diversity, limited resources, and uneven access to native-speaking environments pose persistent challenges, AI offers a way to create adaptive, personalized, and scalable learning systems.

AI-driven technologies such as ICALL (Intelligent Computer-Assisted Language Learning) with Natural Language Processing (NLP) at its core, speech recognition, and adaptive learning algorithms can be strategically applied to address the distinct linguistic and pedagogical needs of Indian learners. It can simulate human-like tutoring by understanding learners' input, diagnosing errors, and providing personalized feedback. For instance, NLP-based systems can analyze common grammatical or syntactic errors while AI-powered speech recognition can help students refine pronunciation and prosody. It is witnessed that these areas are often neglected in large classrooms due to time constraints and lack of native input. Let us see its potential applicability in Korean Language Classrooms:

a. Grammar and Syntax Learning

- Korean has a morphologically rich and agglutinative structure (particles, verb endings).
- NLP-based ICALL tools can segment and analyze word forms to provide feedback on:
  - Proper use of particles (은/는, 이/가, 을/를)
  - Honorifics and speech levels
  - Verb conjugations and sentence structure

b. Pronunciation and Speaking Practice

- Speech recognition systems trained in Korean can help learners:
  - Practice pronunciation and intonation
  - Receive instant, phoneme-level feedback
  - Improve fluency through dialogue simulations

c. Writing and Composition

- NLP-powered writing assistants (like Grammarly but for Korean) can:
  - Detect grammar, spacing, and style errors
  - Suggest natural phrasing based on native Korean usage
  - Support essay and report writing

d. Reading and Vocabulary Acquisition

- Intelligent systems can:
  - Automatically highlight difficult vocabulary and provide contextual meanings
  - Track frequently misunderstood words
  - Recommend readings based on learner proficiency

e. Listening and Comprehension

- ICALL systems with NLP can:
  - Adjust listening exercises based on learner comprehension levels
  - Identify misunderstood phrases and provide explanations

f. Cross-Cultural and Contextual Learning

- Korean ICALL systems can embed cultural elements (e.g., proverbs, honorific use in context)
- Learners can interact with AI tutors or chatbots simulating real Korean communication contexts

So, the advantages of ICALL and NLP in the Indian context are:

- Personalized learning paths for diverse L1 speakers,
- Objective & consistent feedback,
- Bridge between linguistic diversity and unified Korean learning goals,
- Scalable solutions for large university classrooms etc.

Santhosh Viswanathan, Vice President and Managing Director of Intel India Region, explained how AI can make learning more personalised, interactive, and accessible. He noted that India's student-to-computer ratio is around 1:120, compared with 1:1 in many developed countries, and highlighted the role of vernacular AI (Vernacular.ai was an Indian startup founded in 2016 that specialized in AI-powered voice automation and customer service solutions, particularly for enterprise contact centers. They developed the Vernacular Intelligent Voice Assistant

(VIVA) platform to understand and respond to customers in multiple Indian languages and dialects, aiming to improve customer experience and engagement) in bridging gaps in rural and underserved areas.

Beyond individual learning, AI can support instructors through automated assessment, corpus-based error tracking, and intelligent tutoring systems that provide real-time analytics on learner progress. These tools could help teachers design targeted interventions rather than one-size-fits-all lesson plans. Moreover, integrating AI with immersive technologies such as Virtual or Augmented Reality can recreate authentic Korean cultural and linguistic environments, enabling learners to interact contextually in simulated spaces such as markets, offices, or social gatherings.

According to Viswanathan, AI is meant to support teachers by handling repetitive tasks like grading and doubt-solving, allowing educators to focus on mentorship, skill development, and encouraging curiosity among students.
ICALL and NLP are not merely technological tools but pedagogical innovations that can revolutionize how Korean is taught in multilingual environments like India. They enable adaptive, inclusive, and culturally sensitive language learning experiences, helping learners from diverse linguistic backgrounds master Korean efficiently and meaningfully. There are few platforms like Mirinae.io, Memrise, Pingo AI, DuoLingo and 세종학당 AI선생님 which are already taking lead in this direction.

Challenges

For AI to realize its potential, it must be localized and contextually aware. Generic, global AI systems, often trained on English, cannot accurately predict the learning patterns of Indian students whose linguistic backgrounds differ typologically. Developing an India-specific Korean learner corpus, incorporating bilingual datasets (e.g., Hindi–Korean, Tamil–Korean), and training AI systems on these corpora will be crucial steps forward. Collaboration between Indian universities and Korean research institutes can accelerate this process, combining linguistic expertise with technological innovation.

The success of AI integration depends on teacher readiness and ethical design. AI should be positioned as an assistant to the human teacher, not a replacement. Pedagogical decisions such as balancing machine-generated feedback with cultural and communicative nuance must remain under the teacher's control. Ethical considerations, including data privacy, algorithmic transparency, and equitable access, should form the foundation of any AI-driven language education policy.

The Way Forward

For Korean language education in India to thrive in the AI era, technological innovation must go hand in hand with pedagogical wisdom and cultural sensitivity. The goal is not to mechanize learning, but to enhance it, to make language education more responsive to the diversity of Indian learners while preserving the human touch that defines effective communication.

If approached ethically and collaboratively, AI can transform Korean language education from a classroom-bound endeavor into a dynamic, interactive, and inclusive space of intercultural exchange reflecting the very spirit of contemporary India–Korea relations.

References:

Blake R. J. (2008). CALL and its Evaluation, Brave New Digital Classroom. Washington D.C.
Kern R. (2006). Perspectives on Technology in learning and Teaching Languages. Tesol Quarterly.
Government of India. (2001). Census of India. (online) http://censusindia.gov.in.
News18. (n.d.). AI to empower, not replace teachers: Intel India MD on the future of learning. News18. Retrieved [Oct. 10, 2025], from https://www.news18.com/education-career/ai-to-empower-not-replace-teachers-intel-india-md-on-the-future-of-learning-9617477.html
Srivastava, S. (2023). A Study on Development of Korean Language e-Learning System for Indian Learners (Doctoral dissertation, Jawaharlal Nehru University).
UNESCO. (2009). Investing in Cultural Diversity and Intercultural Dialogue. Paris: UNESCO.

POST AI를 향한 한국어교육

Korean Education After AI era

곽용진
(주)이르테크 대표이사

Yongjin Kwak
CEO, IIR TECH

초록

AI를 활용한 한국어 학습(또는 외국어, 제2언어 학습)이 보편적인 수준으로 활용되고 있지만, 여전히 AI가 한국어 또는 외국어 교육(이하 한국어교육으로 지칭하기로 함)에 어떻게 활용되야 하는지에 대해서는 논의가 계속되고 있다. 그러므로, 여기서는 AI와 한국어교육에 대한 논의사항 4가지를 제시하고 향후 AI가 한국어교육에서 어떻게 기여해야 하는지를 살펴보고자 한다.
AI 시대에서 한국어를 비롯한 외국어 교육은 얼마나 많은 노력과 시간이 소요되며, 그 시간과 비용은 AI가 가져오는 편의성과 불확실성에 비해 지불 가치가 있는지를 제기한다. 다른 학습 분야보다 상대적으로 많은 노력과 시간이 소요되는 언어 학습에서 AI가 활용되는 사례들을 통해 언어 학습에 기여하는 부분을 살펴보고, 언어의 총 소요 학습시간 단축을 위한 방안을 간략히 살펴본다. 이를 통해 한국어를 비롯한 외국어 교수-학습에서 AI 자체의 언어 교육적 능력의 발전보다 학습자의 학습활동 데이터의 수집과 분석을 통한 학습시간 단축 방안을 강조한다.

Abstract

AI-assisted Korean language learning (or foreign language or second language learning) is becoming universal method. But, we continues over how AI should be utilized in Korean or foreign language. Therefore, this article presents four key points of discussion regarding AI and Korean language education and explores how AI should contribute to Korean language education in the future.
In the AI era, this article raises the question of how much effort and time is required to teach Korean and other foreign languages, and whether this time and cost are worth the convenience and uncertainty that AI brings. This article examines the contributions of AI to language learning, a field that requires relatively more effort and time than other learning fields. Furthermore, it briefly explores ways to reduce the total learning time required. Through this analysis, it emphasizes the importance of reducing learning time through the collection and analysis of learner learning activity data, rather than the development of AI's inherent language teaching capabilities in Korean and other foreign language teaching and learning.

AI가 가능한가라는 문제는 논제가 되기 어려워졌다는 점에서 AI시대는 이미 도래했고, AI를 통해 하고자 했던 것들은 이제 실현해 보는 문제만 남았다. GPT와 같은 LLM(Large Language Model)의 등장과 함께 AI를 활용하고자 한 가장 활발한 분야 중 하나는 교육이었고, 그 중에서도 외국어 학습은 비교적 실용적이고 널리 활용되는 분야이다.

AI를 활용한 한국어 학습(또는 외국어, 제2언어 학습)이 보편적인 수준으로 활용되고 있지만, 여전히 AI가 한국어 또는 외국어 교육(이하 한국어교육으로 지칭하기로 함)에 어떻게 활용되야 하는지에 대해서는 논의가 계속되고 있다. 그러므로, 여기서는 AI와 한국어교육에 대한 논의사항 4가지를 제시하고 향후 AI가 한국어교육에서 어떻게 기여해야 하는지를 살펴보고자 한다.

1. AI 시대에 외국어를 학습할 필요가 있는가?

AI 기술, 특히 자동 통역 및 번역 기능이 고도화되면서 외국어 학습의 불용성이 대두하고 있다. AI 시대에도 외국어 학습의 필요성은 지적, 사회적, 문화적 가치 측면에서 다음과 같은 면에서 여전히 중요하다고 강조되고 있다.

가. 지적 및 사회적 가치

언어 학습은 자동 통역 기술의 발전으로 활용 가치가 줄어들더라도, 지적 능력으로서의 가치는 남는다. 또한, 여러 언어를 학습함으로써 인간과 인간 사회에 대한 새로운 고찰과 전환기를 열 수도 있다. 특히 언어의 장벽이 민족, 문화, 그리고 소통에 지대한 영향을 미치므로 새로운 언어를 습득하는 노력 자체의 가치는 유효하다고 할 수 있다.

나. 인간 상호작용 및 문화적 맥락의 이해

AI가 메뉴를 번역하거나 호텔에 체크인하는 등 실용적인(utilitarian) 외국어 활용을 대체하더라도 인간 관계를 형성하는 실생활의 의사소통을 완전히 대체하기는 요원하다. 초기 인간 관계의 소통은 불완전한 통번역과 비언어적 수단으로도 형성되지만, 더 많은 관계 형성에는 더 많은 언어적 소통이 요구된다. 그러므로 영화에서와 같은 완전한 통역 장치가 보급되지 않는 한 외국어 학습을 통한 소통 가치가 소멸되는 것은 요원하다. 또한, 완전한 통역 장치가 사람간의 공감, 창의성, 정서적 지원 또는 문화적 맥락을 완전히 재현할 수는 있을지는 아직 확인되지 않았다.

다. AI 기술의 한계

AI 챗봇이나 번역기가 발전했음에도 불구하고, 여전히 문법적 정확성, 문화적 맥락, 섬세한 언어 사용 (예: 존댓말) 등을 완전히 파악하여 교육하기에는 역부족이라는 지적이 많다. 또한, 언어-문화 교차적인 상황에 대한 지식과 편향성 등도 AI 통번역 기술이 외국어 학습의 필요를 대체할 수 없는 것으로 여겨진다.

그럼에도 모국어가 아닌 언어를 배우는 데 드는 시간과 노력에 비해 그 결과가 가치있다는 것을 학습자에게 설득하는 것은 쉽지 않다. 현재 AI의 확산과 성장 추세는 "당신이 배우고 있는 외국어가 쓸모있게 되기보다 AI통역이 먼저 완성되면 어떻할 것인가?"에 대한 불용론에 대응할 마땅한 근거가 없다. 여전히 전세계에서 수많은 사람들이 진학, 취업, 취미로 새로운 언어를 배우고 있고, 그 과정에서 이미 AI가 활용된다. 그러나, 위 질문에 답할 수 없다면 아이러니하게도 AI가 활용될 외국어 학습의 존재가 소멸될 수 있다. 완전한 AI통역 기술이 완성되기까지 10년 이상의 시간이 더 걸린다고 하더라도 대부분이 사람이 새로운 언어를 배워 충분히 소통할 수 있기까지 더 적은 시간이 소요된다고 단언하기 어렵다.

이처럼 다른 분야의 학습과 달리 유독 외국어 학습이 AI발전에 취약한 것은 질문이 포함하고 있는 두 요소, 즉 시간과 노력에 대한 교환가치의 취약함에 있다. 외국어 학습의 가치에 대해 논외로 하더라도 외국어 학습에 필요한 시간과 노력은 여타 분야에 비해 현격히 높은 편이다. 외국어 학습의 필요가 유지되려면 투입되는 시간과 노력을 현저히 낮출 수 있어야 한다. 그러므로, AI시대에 한국어교육은 AI의 발전에 뒤처지지 않게 학습효율을 높이기 위해 AI를 비롯한 기술 활용에 총력을 기울여야 한다.

2. 한국어(외국어)를 배우는 데 필요한 학습시간은 얼마인가?

다른 분야의 학습에 비해 새로운 언어를 배우는 데 필요한 시간은 명확히 산정하기 어렵다. 미국 국무부 산하 외교관 언어 연수 전문 기관인 '외교연구원'(FSI, Foreign Service Institute)은 말하기와 듣기가 일정 수준에 도달하는 데 필요한 시간(주당 수업 시간은 23시간 , 자율 학습 시간은 17시간)을 4개 수준으로 산출했다. 여기서 한국어는 영어를 모어로 하는 학습자에게 88주, 2,200시간의 학습이 필요한 category 4로 산정되었다.

Arabic	Chinese – Cantonese	Chinese – Mandarin
Japanese	Korean	

한국어와 달리 영어 모어 학습자들에게 category 1-3에 속하는 인도-유럽어족은 약 600~1,000시간의 학습이 필요한 것으로 측정되었다.

Category I Languages: 24-30 weeks (552-690 class hours)

Languages close to English.

Danish (24 weeks)	Dutch (24 weeks)	French (30 weeks)
Italian (24 weeks)	Norwegian (24 weeks)	Portuguese (24 weeks)
Romanian (24 weeks)	Spanish (30 weeks)	Swedish (24 weeks)

Category II Languages: Approximately 36 weeks (828 class hours)

German	Haitian Creole	Indonesian

측정값의 정확성과 학습자의 타고난 능력, 이전 언어 학습 경험, 그리고 수업 시간 등 여러 요인에 따라 달라질 수 있음에도 불구하고 긴 학습시간이 대상 언어의 학습 난이도로 여겨진다. 여타 학습이 문명의 발달과 함께 학습시간에 변화를 가져오는 것과 달리, 외국어 학습은 오랜 시간동안 학습 효율의 변화가 크게 없는 듯 하다. 긴 학습 시간의 소요를 유발하는 개별 학습과정과 학습 항목을 식별하고 적정 학습 소요를 산출할 수 있다면 개인 맞춤 학습(Adaptive Learning)을 통해 학습시간을 단축할 수 있을 것이다.
많은 교수법, 학습 자료, 학습 성과에 대한 연구들은 학습 효과와 효율을 향상시켜 외국어 학습 소요시간을 단축하려는 노력이었다고 할 수 있다. 그러므로 효과적인 외국어 학습을 위해 AI가 활용되고 기여하는지 살펴볼 필요가 있다.

3. 외국어 학습에 AI는 어떻게 활용되는가?

AI를 활용한 외국어 학습이 학습기간을 단축하는 명확한 효과와 목표, 성과를 제시하고 있지는 않다. 그러나, 지금까지 제시된 AI활용 방법들을 통해 그 지향하는 바를 살펴볼 수 있다.

가. 개인 맞춤형 학습을 통한 비효율 제거 (Personalization & Adaptivity)

AI는 학습자의 지식 수준, 선호도, 속도, 인지적/정의적 특성을 실시간으로 분석하여 학습 경로와 콘텐츠를 개별화/적응적으로 제공한다.

- 난이도 최적화: AI는 학습자가 너무 쉽거나 어려워 포기하는 상황(조기 포기 가능성)을 줄이고, 수준별 맞춤 학습을 제공함으로써 학습의 지속성과 몰입도를 높인다.
- 자율성 및 동기 증진: 적응형 플랫폼은 학습자에게 자신의 목표를 설정하고 진행 상황을 추적할 수 있는 자율성을 제공하여 내재적 동기를 향상시킨다.

나. 즉각적이고 정교한 피드백 제공 (Immediate and Precise Feedback)

전통적인 교실 환경의 한계인 개별 피드백의 부재 문제를 해소함으로써, 학습자가 오류를 즉시 수정하고 언어를 체화하는 속도를 높여 노력 대비 효율을 높인다.

- 실시간 교정: 자동 쓰기 평가(AWE) 도구는 문법, 철자, 구문 등에 대한 즉각적인 피드백을 제공하여, 학습자가 오류를 바로 인식하고 수정하도록 돕는다. 이는 1:1 맞춤 교육을 실현시켜 1명의 교사가 다수의 학습자를 지도함으로써 발생하는 비효율을 제거한다.
- 발화 능력 평가: AI는 음성 인식 기술을 활용하여 발음, 강세, 억양, 유창성, 정확성 등을 평가하고 수치화된 객관적인 피드백을 제공한다. 말하기에서 이러한 즉각적인 피드백은 학습자가 자기 수정 활동을 통한 반복을 자극해 학습 효율을 높인다.

다. 시공간적 제약에 의한 비효율의 제거

AI 챗봇이나 대화형 에이전트는 학습자가 24시간 언제 어디서나 심리적 부담 없이 대화 연습을 할 수 있는 기회를 제공한다. 교실 수업과 1:다 교육으로 제한된 발화 기회 부족 문제를 완화한다.

라. 평가 자동화

학습자의 수준과 성취 장애를 판별할 수 있도록 평가 문항의 생성, 평가시행, 분석을 자동화함으로써 AI에 의한 개인 맞춤형 학습, 실시간 평가, 피드백을 지원한다. 또한, 평가 자동화는 교실 수업에서 교사의 업무 부담을 완화시켜 수업에서의 학습 효과를 증진하는 데 기여한다.

지금까지 AI를 활용한 외국어 학습은 교사가 학습자에게 교실 수업을 통해 제공하던 것을 모사해 학습을 보조할 뿐 전체 학습 시간-노력의 감소량을 정량적으로 특정하여 제시하는 것으로 보기는 어렵다. 다만, AI가 학습 속도와 성과를 개선하는 데 기여할 수 있음을 실증적 연구로 제시하기도 한다. 황영아 외(2025)는 AI 말하기 앱(플랭)을 활용한 3개월간의 비교과 프로그램 연구 결과, 학습자들은 총 학습 시간 평균은 14시간 18분을 통해 평균 98포인트의 레벨 향상이 이루어져 AI가 실제 말하기 역량 향상에 긍정적 효과를 미쳤음을 보여주며, 학습 지속성이 유지될 경우 누적되는 효과는 시간 단축에 기여할 수 있음을 시사한다.
유사 연구 사례나 시도는 다양하게 나타나지만, 학습자가 목표 언어를 습득하는 과정에서 지속적으로 학습 시간, 학습 활동, 반복, 혼동, 소실, 성취, 숙달 정도가 측정되고 분석된 사례는 발견하지 못했다.

4. 외국어 학습시간은 단축될 수 있는가?

효과적인 외국어 교수 학습법은 지속적으로 연구되고 있으나 획기적 단축 사례는 확인되지 않는 듯 하다. 이는 방법적인 문제보다 아래와 같은 전통적인 교육 환경 문제에서 기인하는 것으로 보는 의견이 많다.

- 제한적인 발화 시간: 다수의 학습자들 대상으로 교육이 수행되므로 학습자 1인당 말하기 시간이 학습자 수에 반비례해 교수 목표 수준에 학습자 전체가 도달하기 어려움.
- 실시간 상호작용의 어려움: 전통적인 교실 환경에서는 실시간 상호작용과 즉각적인 피드백을 제공이 제한되어 학습자가 오류를 인식하고 개선할 기회와 성과가 제한됨.
- 수준별 수업의 한계: 수준별 수업을 시행하고 있으나 같은 교실 내의 학습자가 모두 동일한 수준일 수 없음.
- 주관적 평가와 피드백 부족: 말하기나 발표 과제의 경우, 구체적이고 체계적인 피드백 없이 점수만 부여되는 경우가 많고 평가자의 평가 일관성도 일정하게 유지되기 어려움.
- 불안감 및 자신감 부족: 교실 수업에서는 성취가 낮은 학습자일수록 낮은 자신감 부족으로 발화 활동 자체를 회피하려는 경향이 발생.

앞서 살펴본 AI를 활용 학습들은 이러한 문제점을 개선하는 데 주목하여 새로운 가능성을 보이고 있다. 이러한 교수-학습의 문제 완화가 학습 효과를 높일 것은 자명해 보이나 명확한 학습 소요 기간의 단축 효과를 확인하기는 어렵다. 또한, AI를 활용한 학습 지원은 교사 활동을 모사하여 대체하는 경향을 보이는데, 전통적인 교사-학생 상호작용에는 미치지 못하거나 부정적인 효과가 있다는 의견도 있다(Abbas et al., 2025).

그럼에도 불구하고 AI를 활용한 학습 지원은 학습 시공간의 제약을 극복하고 개인화 학습을 지원함으로써

교사가 학습자 개개인에 대응할 수 없었던 기존 한계를 탈피할 수 있는 기반을 마련했다. 이는 학습자가 언어를 배우는 모든 학습활동을 지속적으로 수집하고 분석할 수 있게 되었음을 의미한다. 총 학습시간은 학습활동*반복수로 산정해 볼 수 있는데, 학습자의 학습 활동을 수집해 반복이 필요없는 학습 활동을 제거하는 만큼 총 학습시간을 감소시킬 수 있다.

그러므로 AI와 디지털 학습 환경이 가져온 변화는 학습자의 상태를 디지털 데이터로 전환해 학습시간 단축을 촉진하는 데 집중할 수 있는 기반을 제공했다고 할 수 있다. 물론 효과적인 학습방법으로 학습활동의 소요시간과 반복수를 함께 줄이거나 제거하는 노력은 여전히 지속되어야 한다. 아직까지 AI기술의 신뢰성과 성능 부족할 뿐만 아니라, 학습 과정에서 불필요한 반복이나 학습활용이 어떤 것이며 얼만큼 차지하는지 충분한 데이터가 확보되지 못했다.

예를 들어, AI의 실시간 평가와 피드백에 의한 학습효과가 평가, 피드백에 의한 직접적인 효과인지 학습자 스스로 문제를 인식하고 개선하는 자극에 의한 반복 효과인지 명확하지 않다. 그러므로 평가와 피드백에 대한 품질과 성능을 지속적으로 개선하는 것과 별개로, 대상 학습활동에 대한 데이터 수집과 상관관계 분석을 학습시간 단축에 대한 노력도 시급하다. AI의 성능향상은 번역과 같은 언어능력의 향상을 수반하지만, 인간의 외국어 학습 비용의 감소는 AI의 성능향상을 통해 직접 얻어지지 않는다.

지금까지 살펴 본 4가지 관점에서 AI 시대의 도래가 한국어를 비롯한 외국어 교육에 위협과 기회를 동시에 제공하고 있음을 제시했다. POST AI 시대의 한국어를 비롯한 외국어 교육은 빠르게 성장하는 AI를 적극적으로 활용해서 AI가 제공하는 통번역과 같은 편익에 학습자의 학습 노력 비용이 매몰되지 않도록 간극을 좁히는 노력이 절실하다.

Abbas, A., Fatima, I., Smavia, S. (2025). A COMPARATIVE STUDY OF TRADITIONAL AND AI SUPPORTED TEACHING METHODS, ENHANCING LANGUAGE LEARNING AND TEACHING THROUGH ARTIFICIAL INTELLIGENCE IN PAKISTANI UNIVERSITIES. INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL) volume 3, 773-792

Zikria, Y. B. (2020). Improving mispronunciation detection of arabic words for non-native learners using deep convolutional neural network features. Electronics, Vol. 9, No. 6, pp. 963. https://doi.org/10.3390/electronics9060963

Baker, T., Smith, L., & Anissa, N. (2019). Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges. Nesta, Retrieved from https://www.nesta.org.uk/report/education-rebooted.

Blyth, C. (2023, June 8). Exploring the A!ordances of AI Tools for L2 Creative Writing [Conference presentation]. CALICO Annual Conference, Minneapolis.

Bong, M., Kim, S., Reeve, J., Lim, H., Lee, U., Jiang, Y., Kim, J., Kim, H., Noh, A., Noh, E., Baek, S., Song, J., Shin, J., Ahn, H., Woo, Y., Won, S., Lee, K., Lee, M., Lee, S. K., & Hwang, A.. (2022 (April 4). SMILES (Student motivation in the learning environment scales): A measure of student motivation in learning environments. Unpublished scale, Brain and Motivation Research Institute, Korea University, https://bmri.korea.ac.kr/

Bonner, E., Lege, R., & Frazier, E. (2023). Large Language Model- Based Artificial Intelligence in the Language Classroom: Practical Ideas for Teaching. Teaching English with Technology, 23(1), 23–41. https://doi.org/10.56297/bkam1691/wieo1749

BotPenguin. (2025 [January 15) Meet Mitsuku: Our virtual friend. An AI chatbot online, https://botpenguin.com/blogs/mitsuku-chatbot

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta- analysis. Language Learning, 67(2), 348–393. https://doi.org/10.1111/lang.12224

Cai, L., & Lee, K. (2021). Exploring the impact of AI-driven grammar correction tools in ESL writing classrooms. Computers & Education, 174, 104295. https://doi.org/10.1016/j.compedu.2021.104295

Cambridge University Press & Assessment. (2024). Generative AI Idea Pack for English language teachers. https://www.cambridge.org/sites/default/files/media/ documents/GenAI_Idea_Pack.pdf

Kan, J. S., Park, M. K., Lee, J. Y., & Lee, Y. E. (2023). The impact of AI-adaptive learning on TOEIC academic achievement and self-directed learning competency in college English classes. The Journal of Learner-Centered Curriculum and Instruction, 23(19), 267–283. https://doi.org/10.22251/jlcci.2023.23.19.267

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., & Krusche, S. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kern, R. (2024). Twenty- first century technologies and language education: Charting a path forward. The Modern Language Journal, 1–19. https://doi.org/10.1111/modl.12924

Kim, D. H. (2023). AI curriculum design for Korea K-12 AI education through analyzing AI education curriculum. International Journal of Recent Technology and Engineering (IJRTE), 12(3), 72-81.

Kim, H. E., Cho, Y. H., & Park, S. H. (2023). Exploring the interaction patterns between learners and AI translator in English writing. Journal of Korean Association of Education Information and Media, 29(1), 201-228. http://dx.doi.org/10.15833/KAFEIAM.29.1.201

Kim, H. J., & Kim, J. R. (2022). An analysis of the research trends of artificial intelligence technology in English education before and after pandemic. Studies in Foreign Language Education, 36(3), 227–242.

Kim, J. (2024). Leading teachers' perspective on teacher-AI collaboration in education. Education and Information Technologies, 29(7), 8693-8724.

Kim, J. A., Kang, D. S., & Ko, Y. C. (2023). A study on educative utilization of generative AI - Focusing on ChatGPT utilization. Journal of the Korean Association of information Education, 27(6), 691-704. http://dx.doi.org/10.14352/jkaie.2023.27.6.691

Kim, J., & Cho, Y. H. (2023). My teammate is AI: Understanding students' perceptions of student-AI collaboration in drawing tasks. Asia Pacific Journal of Education, 1-15. https://doi.org/10.1080/02188791.2023.2286206
Kim, K. L. (2023). A study on self-directed learning ability for EduTech effectiveness. Journal of educational studies, 54(1), 1-22. http://doi.org/10.15854/jes.2023.03.54.1.1
Kim, M. H. (2023). A study on teaching and learning methods for Korean writing using ChatGPT. The Journal of Literacy Creative Writing, 22(2), 55-86. http://doi.org/10.47057/jklcw.2023.58.03
Kim, S., & Bang, J. (2019). One more way of understanding the education in the era of AI. The Journal of Educational Principles, 24(1), 83-105. http://dx.doi.org/10.19118/edp.2019.24.1.83
Kim, S., & Bang, J. (2020). The introduction of the concept of 'Education AI': The challenge of sustainable education by the cooperation between human and AI. The Journal of Educational Principles, 25(1), 1-21. http://dx.doi.org/10.19118/edp.2020.25.1.1
Kim, Y. M. (2023). Exploring the potential application of a conversational AI chatbot in Korean language education - An interaction analysis between advanced learners and ChatGPT -. The Study of Korean Language and Literature, 76, 261-292. http://dx.doi.org/10.15711/WR.76.0.9
Klekovkina, V., & Denié-Higney, L. (2022). Machine translation: Friend or foe in the language class- room?. L2 Journal, 14(1), 105–135. https://doi.org/10.5070/l214151723
Knox, J., Williamson, B., & Bayne, S. (2019). Machine behaviourism: Future visions of 'learnification' and 'datafication' across humans and digital technologies. Learning, Media and Technology, 45(1), 31-45. https://doi.org/10.1080/17439884.2019.1623251
Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal.
Kukulska-Hulme, A. (2020). Mobile-assisted language learning and the role of AI: Trends and future directions. ReCALL, 32(2), 200-217. https://doi.org/10.1017/S095834402000017X
Kundu, A., & Bej, T. (2025). Transforming EFL teaching with AI: A systematic review of empirical studies. Computer Assisted Language Learning, 38(2), 151–174.
Kwak, Y., & Pardos, Z. A. (2024). Bridging large language model disparities: Skill tagging of multilingual educational content. British Journal of Educational Technology, 55(5), 2039-2057. https://doi.org/10.1111/bjet.13465
Kwon, E. Y. (2021). Research trends in AI-based English language teaching and learning. Korea Journal of English Language and Linguistics, 21:1313–1337.
Lan, Y.-J., & Chen, N.-S. (2024). Teachers' agency in the era of LLM and generative AI: Designing pedagogical AI agents. Educational Technology & Society, 27(1), I–XVIII. https://doi.org/10.30191/ ETS.202401_27(1).PP01
Lee, H. (2020). A systematic review of artificial intelligence use in English learning: Focus on higher education. The Journal of Humanities and Social Science, 11(6), 2027–2042. https://doi.org/10.22143/HSS21.11.6.143
Lee, J. H., & Ahn, S. H. (2023). Case analysis for developing key functions of AI digital textbooks. Journal of Creative Information Culture, 9(4), 379-387. http://dx.doi.org/10.32823/jcic.9.4.202311.379
Lee, N. H., & Cha, J. W. (2023). Possibility of generative AI to improve Korean learners' conversational ability. Korean Language, 72, 59-90. http://dx.doi.org/10.52636/KL.72.3
Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. Computer Assisted Language Learning, 35(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323
Liu, S., & Yu, G. (2022). L2 learners' engagement with automated feedback: An eye-tracking study. Language Learning & Technology, 26(2), 78–105. https://www.lltjournal. org/item/10125-73480/
Lu, X., Xu, F., & Han, B. (2022). Artificial Intelligence in Language Learning: Enhancing Personalized Education. Journal of Educational Technology & Society, 25(1), 37-50.
Lv, Z. (2023). Generative Artificial Intelligence in the Metaverse Era. Cognitive Robotics, 3, 208–217. https://doi.org/10.1016/j.cogr.2023.06.001
McWhorter, J. (2023, July 25). Will translation apps make learning foreign languages obsolete? New! York Times. https:// www. nytimes. com/ 2023/ 07/ 25/ opinion/ translation- apps- foreign- languages. html
Ministry of Education of Korea. (2023). Realizing personalized education for all: Strategies for digital-based educational innovation. Ministry of Education of Korea. [Press release] Retrieved from https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=94011&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N.
Noh, Y. (2024). The impact of ChatGPT-assisted learning on English speaking proficiency of Korean university students. The New Studies of English Language & Literature, (89), 89–114. https://doi.org/10.21087/nsell.2024.11.89.89
Park, E. (2024). Study on university English classes using artificial intelligence chatbots. Journal of the Future of Society, 15(1), 231–241.
Park, Y., Lee, S., Hong, S., & Hwang, Y. (2024). A grounded theory on AI-based teacher supporting platform. Journal of Korean Association for Educational Information and Media, 30(3), 1005-1034. http://dx.doi.org/10.15833/KAFEIAM.30.3.1005
Raygan, A., & Moradkhani, S. (2022). Factors influencing technology integration in an EFL context: investigating EFL teachers' attitudes, TPACK level, and educational climate. Computer Assisted Language Learning, 35(8), 1789-1810.
Rha, K., & Baek, J. (2025). Comparing the characteristics of English compositions by Korean EFL college students and AI tools. The Journal of Social Science, 10(1), 455–472. https://doi.org/10.48033/jss.10.1.21
Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. IEEE Transactions on learning and Technologies, 15(4), 481-492. http://doi.org/10.1109/TLT.2022.3196278
Taj, S., & Khan, M. A. (2024). Comparing Grammarly and ChatGPT for automated writing evaluation of ESL learners. Bahria University Journal of Humanities and Social Sciences, 7(2), 63–90.
Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., & Pange, J. (2023). Exploring the potential impact of artificial intelligence (AI) on international students in higher education: Generative AI, Chatbots, analytics, and international student success. Applied Sciences, 13(11), 6716.
Xiu-Yi, W. (2024). AI in L2 learning: A meta-analysis of contextual, instructional, and social-emotional moderators. Language Learning & Technology, 28(1), 1–25.

AI 시대의 영어교육(TEFL)에서 돌봄 중심 교수법

Care-Based Pedagogy in TEFL the AI Age

조지은
옥스퍼드대학교 교수

Jieun Joe Kiaer
Professor, University of Oxford

Abstract

In this talk, I explore care-based pedagogy as a crucial foundation for sustainable and innovative ELT practices in the AI age. As artificial intelligence (AI) continues to transform education, English Language Teaching (ELT) must evolve to balance technological advancements with the human-cantered elements of teaching. By integrating care, innovation, and sustainability, educators can create learning environments that foster emotional engagement, ethical teaching, and long-term success in language acquisition.

Care in ELT extends beyond academic instruction; it involves empathy, trust, and relational ethics, ensuring that students feel supported, motivated, and valued. In AI-enhanced classrooms, where automation may depersonalize learning, teachers must act as mentors and facilitators, bridging the gap between AI efficiency and human connection. The MERGE framework (Monitor, Encourage, Reward, Guide, Evaluate) highlights the evolving role of educators in an AI-integrated setting, emphasizing their responsibility in shaping both cognitive and emotional learning experiences.

This talk also examines innovation in ELT, particularly the role of AI in adaptive learning, gamification, and personalized language instruction. While AI can provide individualized feedback and enhance engagement, human teachers remain irreplaceable in fostering resilience, creativity, and ethical awareness in learners. Case studies from Asian TEFL contexts illustrate how care-based strategies encourage active participation, especially among students who may be less expressive but excel in structured, supportive environments.

Finally, sustainability in ELT is explored through the lens of long-term pedagogical impact. A care-cantered approach ensures that learning is not only effective but also meaningful, preparing students for lifelong language use rather than short-term test performance. By focusing on care-driven methodologies, educators can contribute to a more inclusive, equitable, and sustainable future for TEFL in an era of rapid technological change.

This talk argues that the future of TEFL lies in integrating AI-driven innovation with the human power of care—ensuring that technology supports, rather than replaces, the essential relational aspects of teaching and learning.

Introduction: From Knowledge Transmission to Care-Based Orchestration

Artificial intelligence (AI) is transforming the ecology of education. In recent years, schools and universities have rapidly adopted machine-based systems capable of analysing data, generating content, and tailoring learning experiences. The OECD (2023) defines AI as "a machine-based system that, for explicit or implicit objectives, infers from input how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." AI now operates on both sides of the classroom. For pupils, adaptive learning platforms, chatbots, and virtual environments deliver personalised content. For teachers, AI assists lesson planning, assessment, and administration. Yet despite its promise of efficiency, AI raises fundamental questions about the purpose of education.

Media discourse often frames AI as a threat—a replacement for teachers or a challenge to academic integrity. This framing obscures a more profound transformation: AI is not eliminating teachers but redefining their professional identity. Teaching is shifting from the transmission of knowledge to the orchestration of learning environments, requiring teachers to balance technology's power with ethical, emotional, and relational intelligence.

At the centre of this transformation lies care. Care—understood as attention to the learner's cognitive, emotional, and moral development—emerges as the foundation of professional competence. This paper argues that *care-based pedagogy* offers a sustainable path forward in the AI age, ensuring that education remains human-centred, ethical, and inclusive.

1. The Evolving Role of the Teacher

The integration of AI does not simplify education; it adds new layers of complexity. Teachers must now make judgements across three intertwined domains:

1. Cognitive orchestration – determining when and how AI supports rather than substitutes human learning.
2. Ethical stewardship – managing issues of authorship, bias, privacy, and wellbeing.
3. Affective care – sustaining motivation, belonging, and trust in technologically mediated learning.

UNESCO (2023) and the OECD (2023) highlight that AI can improve learning outcomes only when human oversight remains central. Without ethical and emotional anchoring, AI amplifies existing inequalities and erodes autonomy.

Teachers thus transition from being *sources* of knowledge to *curators* and *interpreters* of knowledge ecosystems. Their role is not to compete with machines in content delivery, but to guide learners through complex digital environments with discernment and empathy.

This marks a paradigmatic shift from the "sage on the stage" to what Kiaer (2023) calls the care-based orchestrator—a teacher who harmonises human and artificial intelligence for holistic learning.

2. Global Context and Comparative Perspectives

AI adoption in education differs markedly across national contexts.

In the United Kingdom, caution prevails. The Department for Education's 2023 statement reflects uncertainty and concern over integrity, bias, and the ethics of AI-generated work. Many schools have banned AI tools outright, citing risks of plagiarism and data misuse. Most teachers report limited training or institutional guidance, and anxiety about surveillance and workload remains high.

In South Korea, by contrast, experimentation is more systematic. Korea's Ministry of Education piloted AI Digital Textbooks (AIDTs) across multiple grade levels. While a national rollout planned for 2025 is currently paused following political changes, grassroots teacher-led innovation continues. These communities demonstrate a key insight: AI integration succeeds when driven by empowered teachers, not imposed from above.

Teachers in Korea have built vibrant communities of practice—notably via KakaoTalk groups with thousands of members—where they share lesson plans, discuss ethics, and provide emotional support. Participation counts as professional development, acknowledging the emotional and social labour of innovation.

This contrasts with the UK, where professional isolation and policy ambivalence limit experimentation. The comparison reveals that systemic support, recognition, and community care are essential for sustainable AI integration.

3. Emerging Research Questions

The rapid incorporation of AI raises pressing research questions across cognitive, ethical, and evaluative dimensions:

1. Cognitive Load and Authorship
   Cognitive offloading theory (Risko & Gilbert, 2016) suggests that reliance on external tools reshapes attention and memory. Generative AI can undermine metacognitive confidence, reducing active reasoning (Zhai et al., 2024). Teachers must therefore scaffold reflection and self-regulation to maintain deep learning.
2. Safeguards and Ethics
   As AI systems process student data, privacy and authorship become critical. OECD (2023) frameworks emphasise governance and human oversight, yet ethical training for teachers remains patchy. A care-based approach reframes ethics not as compliance but as relational responsibility.
3. Evolving Evaluation
   With AI capable of completing essays and coding tasks, traditional assessment models falter. Kiaer (2023, 2024) argues for new forms of insight assessment—in-class tasks, oral presentations, and process-based evaluations—that value creativity, empathy, and ethical reasoning over rote output.

These questions reveal a central truth: AI challenges not only what students learn, but how teachers define learning itself.

4. Teacher Competencies in the AI Age

The competencies required of teachers in AI-rich environments extend beyond technical literacy. The MERGE framework (Kiaer & Jeon, 2024) provides a model for care-based pedagogy built around five principles:

- Monitor: Use AI analytics judiciously, cross-checking algorithmic data with human observation.
- Encourage: Combine AI scaffolding with empathy and emotional support.
- Reward: Humanise gamified progress through genuine teacher praise and relational feedback.
- Guide: Model ethical AI use—prompting responsibly, crediting sources, and identifying bias.
- Evaluate: Assess learning holistically, combining academic attainment with social–emotional growth.

This framework positions care as both ethical compass and pedagogical method. It recognises that emotional attunement, ethical judgement, and contextual understanding remain uniquely human capacities—ones that AI cannot replicate.

5. Lessons from Korea: The Human Infrastructure of AI

Field studies in Korean pilot schools demonstrate how systemic and social structures shape the success of AI adoption.

1. Workload and Recognition
   Korean teachers receive training within paid working hours, integrating AI professional development into institutional routines. This prevents "AI fatigue" and fosters sustained engagement.

2. Competency Gaps and Support Structures
   Kim and Kwon (2023) mapped 22 AI-related teacher competencies under the TPACK model. Teachers showed the lowest confidence in AI content knowledge, coding, and ethics. However, targeted workshops and peer mentoring significantly improved confidence and reduced anxiety.
3. Communities of Practice
   Online teacher networks—particularly the 3,000-member KakaoTalk AIDT group—enable teachers to exchange ideas and emotional support. Participation is recognised as professional growth, reinforcing the link between collaboration and wellbeing.
4. AI and Social–Emotional Learning (SEL)
   AI tools can enhance SEL by offering low-stakes communication platforms. Avatars and chatbots provide safe entry points for anxious students, while AI diary analysis allows teachers to detect stress signals early (Kiaer, 2023). AI thus becomes a mirror for care, not a substitute for it.

The Korean example underscores that AI succeeds when it augments human networks, not when it replaces them.

6. Care-Based Pedagogy: Theoretical Grounding

Care-based pedagogy draws on feminist ethics of care (Noddings, 2012) and humanistic education traditions. It emphasises relational interdependence and emotional reciprocity. In the AI era, these values acquire new urgency.

AI cannot feel empathy or assume moral responsibility. Its recommendations lack contextual nuance and emotional resonance. Teachers' roles, therefore, expand as ethical mediators who ensure that technological efficiency does not override human flourishing.

Care-based pedagogy involves:

- Attentional care – noticing when students disengage or rely excessively on AI tools.
- Reflective care – guiding metacognitive awareness of AI's role in learning.
- Ethical care – modelling transparency, citation, and digital ethics.
- Communal care – building safe learning communities where trust and dialogue counter algorithmic opacity.

Care thus becomes both an epistemic and emotional strategy: it nurtures the conditions for meaningful, responsible learning.

7. Implications for Policy and Practice

1. Teacher Training and Certification
   AI literacy should form part of professional qualification, encompassing not only technical skills but ethical reasoning and SEL integration.
2. Institutional Recognition
   Teachers' emotional labour in navigating AI complexity must be acknowledged in workload models and promotion criteria.
3. Cross-Cultural Learning
   Countries can learn from Korea's community-based model, embedding teacher voice and care networks into AI governance.
4. Ethical AI Design
   Developers must collaborate with educators to ensure transparency, bias mitigation, and human-centred functionality.
5. Student Wellbeing and Agency
   Policies should prioritise reflective engagement over efficiency metrics, cultivating critical, compassionate AI users.

8. Beyond Replacement: Reimagining Teaching as Ethical Mediation

The central anxiety around AI in education stems from a false binary: human versus machine. Yet teaching has always been a relational act, mediated through tools—from chalkboards to chatbots. The task ahead is not resistance to AI, but reimagining teaching as ethical mediation—a space where care bridges human and artificial intelligence.

Teachers, as ethical mediators, model discernment: they demonstrate that wisdom lies not in constant connectivity but in choosing when to pause, reflect, and care. This reorientation is vital for student wellbeing in an age of algorithmic acceleration.

Conclusion: Care as the Core Competence

In the AI age, effective teaching cannot be measured solely by digital proficiency. What defines the teacher is *judgement grounded in care*.

Care is not a sentimental virtue but a professional necessity—anchoring ethical practice, cognitive engagement, and emotional wellbeing. As AI expands, so too must our capacity to humanise it.

Korea's AIDT experience shows that when teachers are supported, recognised, and connected, AI becomes a tool for empowerment rather than alienation. The future of pedagogy will depend on educators' ability to orchestrate—not compete with—technology, turning classrooms into spaces of balanced, compassionate innovation.

Care-based pedagogy, therefore, is not a nostalgic return to pre-digital ideals but a progressive model for sustainable, humane education in the 21st century.

Selected References
Kiaer, J. (2023). The future of syntax: Asian perspectives in an AI age. Bloomsbury Academic.
Kiaer, J. (2024). Conversing in the metaverse: The embodied future of online communication. Bloomsbury Academic.
Kiaer, J., & Jeon, M. (2024). Humanizing AI education: The MERGE framework for supporting teachers in AI-enhanced classrooms. International Journal of Contents, 20(3), 1–8.
OECD. (2023). OECD framework for the classification of AI systems. OECD Publishing.
Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688.
UNESCO. (2023). Guidance for generative AI in education and research. Paris: UNESCO.
Zhai, C., Wibowo, S. & Li, L.D. (2024) The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. Smart Learn. Environ. 11, 28.
Wang, J., Fan, W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. Humanit Soc Sci Commun 12, 621 (2025).

분과회의 세션 12 Parallel Session 12

AI와 언어 AI and Language

인공적 친밀성(Artificial Intimacy)과 인간 관계의 변화

Artificial Intimacy: The Transformation of Human Relationships

이유미
중앙대학교 교수

Yumi Yi
Professor, Chung-Ang University

초록

4차 산업혁명과 팬데믹은 인간관계의 양식을 근본적으로 변화시켰으며, 이에 따라 인간은 비인간적 주체와도 친밀감을 형성하는 새로운 관계 경험을 하고 있다. 본 논문은 '인공적 친밀성(Artificial Intimacy)' 개념을 이해하기 위해서 인간의 친밀성 형성 요소를 분석하고 이를 바탕으로 대화형 AI, 아바타, 로봇 등과의 친밀감에 대한 연구 키워드인 엘리자 효과, Darwinian button, 감정 투사 등을 이해한다. 이러한 이론을 통해 이해할 수 있는 것은 인가은 인간과의 관계 형성에서 나타나는 불편함을 최소화하고 공감에 대한 본능을 충족할 수 있는 AI에 인공적 친밀감을 느끼고 있음을 분석하였다. 더불어, 이러한 사회적 현상을 통해 앞으로의 인간관계와 사회 변화에 대한 바를 방향성을 위해 관계리터러시에 대한 관심과 교육을 주장하였다.

Abstract

The Fourth Industrial Revolution and the COVID-19 pandemic have fundamentally transformed the modes of human relationships, leading individuals to form new types of intimacy with non-human agents. This paper explores the concept of Artificial Intimacy by first analyzing the core elements of human intimacy and then applying this understanding to interactions with conversational AI, avatars, and robots. Key theoretical concepts such as the Eliza effect, Darwinian buttons, and emotional projection are examined to explain how individuals develop artificial intimacy with AI. The study suggests that humans tend to project intimacy onto AI as a means of minimizing the discomfort of human relationships and fulfilling their innate desire for empathy. In addition, the study highlights the importance of developing relational literacy as a means to effectively respond to the evolving nature of human relationships.

1. 기술과 인간관계의 변화

인간은 항상 관계를 고민한다. 엄마의 탯줄로부터 연결되어서 생명을 얻었고, 태어나면서부터 가족이라는 사회 속에서 존재하기 때문에 연결에 대한 고민은 본능일 것이다. 그러나 COVID 19라는 전 세계적인 재난 때문이 아닐지라도 기술의 급격한 발달은 매우 색다른 형태의 연결과 관계를 만들었다. 늘 연결되어 있으나, 누구와도 완전하게 연결되어 있지 못한 관계의 모습이 그것이다.

인간의 편의성을 증대시켜 온 기계문명의 발달은 인간의 욕구를 충족해 오는 과정이기도 하였다. 1차 산업혁명은 농업의 기계화를 통한 인간 생존 욕구를 충족시켰다면, 2차 산업혁명은 전기 에너지 기반의 대량 생산 혁명을 통해 자본이나 사회적 구조에 있어 안정성에 대한 욕구를 충족시켰고, 이후 네트워크 혁명인 3차 산업혁명은 인터넷 혁명을 통해 공간적 한계를 넘어선 네트워크를 가능하게 함으로써 인간의 소속욕구를 충족시켰다. 현재의 4차 산업혁명은 네트워크의 새로운 측면을 고려하여 인간의 자기실현 욕구를 충족하는 방향으로 발달해 가고 있다.

3차 산업혁명 시대에 이뤄진 네트워크의 발달로 인해 인간은 시간과 공간의 제약 없이 어디에 있는 누구와도 연결될 수 있는 자유를 얻었다. 그러나 식탁이라는 물리적 공간 안에 있는 가족들은 또다른 네트워크 안에 있는 타인과 연결되는 것을 통하여 실재 공간의 가족과는 완전하게 연결되지 않는 부작용도 낳았다. 또한, AI 기술의 발달을 통한 관계의 확장은 다양한 기기를 통해서도 연결의 욕구를 충족할 수 있다는 가능성을 보여주었다. AI 스피커나 챗봇과 같은 기계와 연결되는 것 외에도, 아바타를 이용해 선별적으로 자기를 노출하는 맥락에서도 관계 맺음이 가능해졌기 때문이다. 이는 인간관계에서 상대적으로 자존감을 위축시키는 요소를 스스로 제어할 수 있는 환경을 만들어 준 것이라 할 수 있다.

이처럼 기술의 발달로 인한 관계의 변화는 연결되고 싶으나 독립적이고 싶은 인간의 변증법적 욕구를 스스로 온전히 제어할 수 있는 기술을 제공했다는 점에서 의미가 있다. 또한, 스스로 로그인-아웃 여부와 시간을 결정함으로써 인간은 인간이 누군가와 연결되고 싶으나, 독립적이고 싶은 상반된 욕구를 충족한다. 이러한 네트워크 시스템의 속성은 기존의 대인 관계에서 타인과 협상해야 하는 스트레스를 없애 주기에 매우 유용하게 느껴진다.

그러나 이러한 관계가 진정 행복할까? 이러한 근원적인 질문을 던져본다면 현대 사회를 거스르는 우문일까? AI 기술이 만들어 낸 가상 세계의 공간은 현실 세계를 분리한 공간이기보다는 현실의 확장된 공간이라는 점을 생각해 볼 때 결국 인간관계는 다시 현실과 협상하는 단계로 돌아와야 한다. 이러한 측면에서 현대 사회의 인간관계를 균형적으로 살펴보지 않는다면 현실과 가상 세계의 간격은 더 크게 느껴질 것이다.

현대 사회의 인간관계는 과거의 면대면 커뮤니케이션만을 통해서만 형성되지 않는다. 인간의 관계 형성의 문제뿐 아니라 관계의 '대상' 또한 다양하기 때문이다. 우리는 여전히 직접 만나 관계를 맺지만, 매개된 미디어를 통해 더 많은 소통을 이어간다. SNS나 메일 등을 통해 만날 시간과 만남의 의미 등을 확인하고, 장소를 공유하고 나서야 만날 수 있다. 만남이 이뤄진 후에는 다시 SNS나 메일 등을 통해 만남의 후기를 공유하면서 관계를 공고히 해 나간다. 이러한 관계는 그래도 인간과 인간의 관계를 전제한다. 그러나 요즘 사람들은 인간과만 커뮤니케이션 하지 않는다. 매일 아침 '헤이 카카오' '시리야' '빅스비' 등의 호출을 통해 날씨를 확인하고, '엄마에게 전화해'를 외치며 사람이 아닌 누군가에게 명령을 한다. 이제 더이상 인공지능 어시스

턴트라 불리는 기기들과의 대화가 낯설지 않으며, 노안이 있는 어르신들은 오리려 인공지능과 대화를 더 자주 유용하게 시도하고 있다.

이제는 음식점에 들어서면 점원이 우리를 맞이하기 전에 키오스크가 무엇을 먹을지를 물어보고, 서빙 로봇이 음식을 가져다준다. 인간이 아닌 인공지능 시스템을 통해 메시지를 전달하고 메시지 없는 그들의 서빙을 받으면서 편의를 영위하고 있는 것이다. 이와 같은 삶의 변화로 인해 케어로봇에 대한 인식은 점차 긍정적으로 변화했는데, 이러한 변화는 커뮤니케이션 대상의 확장을 반증한다

인간의 관계 형태와 커뮤니케이션 방식이 다양해지는 것은 하나의 현상이다. 이러한 현상을 수용할지의 여부와 현상이 유지될 것인지의 여부를 예측하는 강력한 변수 중 하나는 인간이 경험하는 감정일 것이다. 소원한 관계에 대한 두려움, 외로움, 편안함 등의 감정, 관계 맺기에 개입하는 기술에 대한 즐거움, 호기심, 불편함 등의 감정 등은 매개된 커뮤니케이션의 결과를 예측한다. 감정은 사람 또는 사람이 아닌 주체와 소통하며 경험하는 다양한 느낌들로, 감정을 유발한 대상(관계 또는 커뮤니케이션 방식 등)에 대한 주관적인 인식과도 같다. 감정은 중요한 정보와 맥락을 제공하고, 시간의 흐름에 걸쳐 상호작용이 전개되는 방식에도 영향을 미치게 되므로(Hareli & Rafaeli, 2008), 감정으로 인해 소통 대상자들은 서로를 더 가깝게 느끼기도 하며, 소통의 질, 소통에 대한 태도나 관여 수준, 소통 상대와의 관계 발전 등이 달라지기도 한다.

현대 사회는 미디어로 매개된 커뮤니케이션을 이용해서 인간의 관계를 확장하던 시간을 지나 인간과 로봇의 원활한 소통을 기대하고 있다. 그러므로 지금의 커뮤니케이션을 변화를 살펴보는 것은 인간 커뮤니케이션의 미래를 예측하게 할 것이다.

셰리터클은 현대인이 점점 더 외로워진다고 했다. 이는 어쩌면 현실 세계에서 느끼는 관계의 어려움을 기술 발전이 해결하리라는 기대에서 비롯된 현상일지도 모른다. 그러나 기술을 통해 확장된 세계는 인간 관계를 더욱 복잡하게 만들었기에 이러한 환경에서 나를 행복하게 할 관계의 모습은 어떤 것인지에 대하여 더 많이 고민하고 학습해야만 한다. 인공적 친밀성(Artificial Intimacy)는 현대 사회의 변화된 커뮤니케이션 양상으로 인하여 등장한 용어이다.인간이 아닌 인공에게 느끼는 친밀성을 통해 인간관계에 미칠 영향성을 고민하는 것은 현재뿐 아니라 미래 사회의 관계성을 진단하는 데 있어 중요한 지표가 될 것이다.

2. Artificial Intimacy

1) 인공적 친밀성의 정의

Artificial Intimacy 인공적 친밀성을 정의하는 것은 아직은 완전하지 않다. 인공적 친밀성이 인간이 아닌 인공의 대상에게 친밀함을 느끼는 것이라고만 정의한다면 그 자체로 완전하지만 인공의 친밀성을 느끼는 대상, 이유, 정도 등의 관점에서 이를 확인한다면 간단한 문제는 아니다. 인간의 관계를 파고드는 커뮤니케이션 미디어로 인하여 인간 관계의 변화를 연구해온 Sherry Turkle의 책에서 논의했던 기계와 맺는 친밀성의 환상은 인공적 친밀성을 논의한 시작일 수 있다.

대화형 AI의 발달은 이러한 기계와 맺는 친밀성의 양상을 더욱 심화 발전시켰고 이에 따라 인공적 친밀성이라는 용어가 중요하게 인식되기 시작하였다. 그 한 예로 Aspen Institute는 연례 AI 라운드 테이블 중 네 번째 회의에서 Artificial Intimacy을 주제로 선정한 것이다. 그 보고서(2020)에서 논의된 결과는 인공적 친밀성을 완전하게 정의하지는 못했지만 인간-기계의 상호작용을 탐색의 측면에서 사전적 정의가 아닌 합의된 용어인 작동적 어휘(operational lexicon)의 필요성을 설명하였다. 더불어 이 용어가 사회적 파급력이 큰 주제임을 확인하였고, 이를 통한 연구의 확장은 사회 전반에 중요한 영향성을 미칠 것이라는 결론을 도출하였다.

인공적 친밀성에 대한 중요성 인식은 단순히 대화형 AI 기술의 발전만의 문제는 아니다. 'Artificial Intimacy: Virtual Friends, Digital Lovers, Algorithmic Matchmakers (Brooks,2021)'라는 제목의 책에서는 COVID19 팬데믹이 이러한 인공의 친밀성을 강화시켰다고 보고 있다. 아직은 정의에 있어서도, 그 양상과 특징에 있어 다양한 논의를 이끌고 있지는 않지만, 기술이 인간의 관계를 매개하던 시대에서 대화의 타자 즉 대상이 된 시대로 변화된 시점에서 인공적 친밀성은 관계의 변화를 이해하기 위한 중요한 개념이 될 것이다.

2) 친밀성에 영향을 미치는 요소

친밀성 영향 요소	세부내용	출처
자기노출 (Self-disclosure)	개인의 내면적 정보, 감정, 경험을 자발적으로 타인에게 공유함으로써 신뢰와 유대를 형성함. 자기노출은 친밀감 형성의 핵심 조건이며, 정서적 존재의 표현.	Reis & Shaver (1988), Andreescu (2020), Jamieson (2005)
신뢰 (Trust)	친밀한 관계는 서로를 신뢰할 수 있다는 안정감 위에 형성됨. 신뢰는 자기 개방과 감정 공유의 핵심 기반.	Reis & Shaver (1988), Popovic (2005)
감정적 의사소통 (Emotional communication)	감정을 명확히 전달하고 수용하는 능력은 친밀감 형성에 핵심. 현대 친밀성은 감정 표현 중심임.	Giddens (1992), Popovic (2005), Andreescu (2020)
성적 친밀감 (Sexual intimacy)	성적 개방성과 상호 존중을 바탕으로 신체적·감정적 유대를 형성.	Popovic (2005), Sexton & Sexton (1982)
개인 경계 설정 (Boundary work)	자율성과 정체성을 보호하기 위해 관계 내 경계를 협상하는 과정. 디지털 환경에서는 경계가 유동적임.	Jamieson (2005), Andreescu (2020), Zurbriggen et al. (2015)
자율성과 독립성 (Autonomy & individuation)	건강한 친밀감은 각자의 독립성을 보장할 때 가능함. 과도한 통제는 친밀감 저해.	Popovic (2005), Andreescu (2020)
공감과 이해 (Empathy & understanding)	상대방의 감정·입장을 수용하고 이해하는 태도는 정서적 유대를 강화함.	Reis & Shaver (1988), Popovic (2005)
갈등 시 의사소통 (Communication during conflict)	갈등 상황에서 문제해결 중심, 감정적으로 안정된 커뮤니케이션이 친밀감 유지에 기여.	Overall & McNulty (2017)
사회적 가시성과 피드백 (Social media visibility& feedback)	SNS의 공개성은 친밀감 형성을 촉진할 수도, 위협할 수도 있음. 감정과 관계가 '보여주는 것'으로 환원됨.	Zurbriggen et al. (2015), Andreescu (2020)
성역할 및 성차 (Gender roles &differences)	사회적으로 규정된 성역할이 친밀성 표현에 영향을 미침. 여성은 감정 표현, 남성은 억제의 역할 강조.	Jamieson (2005), Popovic (2005), Andreescu (2020)
감정 표현의 자유 (Freedom of emotional expression)	감정을 자유롭게 표현할 수 있어야 깊은 친밀감 형성 가능. 억압된 관계는 친밀감이 약화됨.	Popovic (2005), Reis & Shaver (1988)
공통 가치와 이상 (Shared values & ideals)	신념·이상 공유는 관계 지속성과 친밀감을 강화하는 기반.	Popovic (2005), Sexton & Sexton (1982)

시간과 관심의 투자 (Time & attention investment)	신념·이상 공유는 관계 지속성과 친밀감을 강화하는 기반.	Reis & Shaver (1988), Popovic (2005)
심리적 안정감 (Psychological safety)	꾸준한 시간과 정서적 관심은 친밀감 유지의 실질적 기초.	Popovic (2005), Reis & Shaver (1988)
프라이버시 통제 및 불안정 (Privacy control & turbulence)	정보 공유의 경계가 불분명할 경우 친밀감은 위협받음. SNS 환경에서 더욱 불안정성 증가.	Zurbriggen et al. (2015), Petronio (2002)
상호의존성과 상호작용 (Mutual dependence & interaction)	의미 있는 상호작용은 정서적 유대를 강화하며 관계의 지속을 돕는다.	Popovic (2005), Reis & Shaver (1988)
의도적 관계 유지 노력 (Intentional relationship work)	친밀한 관계는 의식적인 노력과 감정적 노동을 통해 유지됨.	Giddens (1992), Andreescu (2020)

3) AI와의 친밀성은 왜 형성되는가?

인공적인 대상에 대한 친밀감을 느끼는 대표적인 초기 이론으로는 Eliza 효과가 있다. Eliza는 컴퓨터를 활용한 자연어 대화 프로그램으로 MIT가 개발한 것이다. 이 프로그램은 키워드를 기반으로 대답을 생성하는 고전적인 대화 프로그램이다. 대화 상대자의 말을 단순 반복적인 형태로 응대함에도 불구하고 상호 이해의 환상(illusion of mutual understanding)으로 사용자는 생각한다고 분석하면서 엘리자 효과를 논의하기 시작하였다.(Weizenbaum,1966, Shrager,2024,Shum et al,2018) 이러한 엘리자 효과는 현대의 AI 대화 시스템에 대한 인공적 친밀감을 설명하는 데 중요한 기초가 된다. 엘리자 효과의 핵심은 AI라는 인공 시스템에 대한 의인화 즉, 이는 사용자가 자신의 생각과 감정을 프로그램에 투사함을 통해 생성된다. 이러한 투사의 가장 핵심은 공감이다. 인공적 친밀성에 대한 보고서를 낸 Aspen Institute Report(2020)에서도 Darwinian Buttons이 가진 친밀성의 요인을 제시한다. 이 용어는 셰리 터클을 통해서 설명되는데, 이는 인간이 인공의 대상으로부터 공감받는다고 느끼게 하는 것을 의미한다. 친밀한 형태는 감정의 투사를 통한 공감을 스스로 조작하고, 친절한 말투는 상대가 나를 이해하는 듯이 느끼게 하는데 이는 상대의 진실성과 의도에 관계 없는 거짓 공감(pretend-empathy)이며 이것을 결국 인간 스스로 공감을 착각하는 행동인 것이다.

이 모든 심리적인 인공적 친밀성은 단순히 앞서 본 인간 간의 친밀성의 요소를 인공의 대상으로부터 느끼는 것일 수도 있는데 이것의 핵심은 갈등하지 않은 상대이며, 상대는 나에게 호의적인 상태라고 생각하는 것에서 출발한다. 세리 터클이 말했듯, 인간은 상대가 자신을 모른다고 보기에 무엇을 문제인지를 묻는 것으로부터 나에 대한 공감은 출발하게 되는데, 이러한 작은 시작을 AI와 함께 하게 되면 그 다음의 과정에서 인간과 다른 친밀함의 과정을 경험하게 되는 것이다. Aspen(2020) 보고서나 Brooks(2021) 저서에서 이야기 하고 있는 이러한 인공적 친밀함 확산의 중요한 요인으로 고립을 강제 경험하게 된 팬데믹으로 보고 있으며, 인간의 필요와 인간으로부터의 독립을 모두 충족하는 방식으로 AI를 택한 것이라 하겠다.

3. 관계 리터러시

대화형 인공지능의 발달은 커뮤니케이션 매체의 위상을 새롭게 정의하였다. CMC라는 용어가 정의하듯이 커뮤니케이션 매체는 인간과 인간을 매개하는 대상이었지만, 대화형 인공지능의 등장은 인공지능이 매개체가 아니라 인간 커뮤니케이션이 타자 즉 대상이 되었다. 이러한 사회적 변화는 인간 커뮤니케이션 환경에 따라 변화하는 인간성의 변화, 이를 통한 사회의 변화라는 측면에서 인공지능이 인간관계와 사회를 어떻게 변화시키게 될 것인가는 중요한 문제가 되었다. 인공지능의 발달은 관계에 따른 갈등 관계를 선택적으로 두고자 하는 욕망을 크게 하였고, 관계에 대한 불편함을 최소화하고자 하는 인간 사회의 현상을 만들기도 하였다.

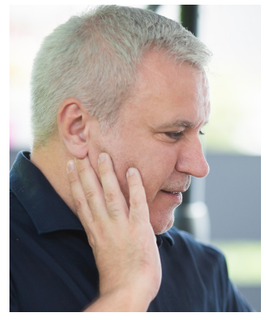
그러나 확실한 부분은 인간은 인간 속에서 사회를 형성하고 이를 통해 발전해 가기 때문에 인간 관계른 불가피한 부분이라는 점에서 새롭게 정의되어가는 인간 관계 현상을 객관적으로 분석할 수 있어야 할 뿐 아니라, 새로운 세대에 맞는 그리고 만들어가야 하는 형태의 인간관계를 위한 새로운 리터러시를 개발해야 할 것이다. 관계는 배우는 것이 아니라 성장하는 것이라는 관점에서 관계 또한 사회 구성원으로 살아가기 위해 배워야 하는 대상이라는 관점을 가질 필요가 있는 것이다. 이러한 관점에서 새로운 AI 시대에 인공적 친밀성을 중요한 화두로 둘 뿐 아니라 이로 인한 다양한 문제를 해소하기 위한 관계 리터러시에 관심을 가져야 할 때이다.

<참고문헌>

Andreescu, F. C. (2020). Rethinking intimacy in psychosocial sciences. Psycho-Politics International, 18(1), 22–38.
Brooks, R. (2021). Artificial intimacy: Virtual friends, digital lovers, algorithmic matchmakers. Columbia University Press.
Giddens, A. (1992). The transformation of intimacy: Sexuality, love and eroticism in modern societies. Stanford University Press.
Gloria, K. (2020). Artificial intimacy: Roundtable on artificial intelligence 2020 rapporteur's report. Aspen Institute.
https://www.aspeninstitute.org/publications/artificial-intimacy
Jamieson, L. (2005). Boundaries of intimacy. In M. Knapp & J. Daly (Eds.), Handbook of interpersonal communication (pp. 1–16). Springer.
Overall, N. C., & McNulty, J. K. (2017). What type of communication during conflict is beneficial for intimate relationships?. Current Opinion in Psychology, 13, 1–5.
Petronio, S. (2002). Boundaries of privacy: Dialectics of disclosure. SUNY Press.
Popovic, M. (2005). Understanding emotional intimacy. Psychotherapy: Theory, Research, Practice, Training, 42(4), 432–446.
Reis, H. T., & Shaver, P. (1988). Intimacy as an interpersonal process. In S. Duck (Ed.), Handbook of personal relationships (pp. 367–389). Wiley.
Sexton, C. W., & Sexton, T. L. (1982). The intimate couple: Development, dynamics, and treatment. Prentice-Hall.
Shrager, J. (2024). ELIZA reinterpreted: The world's first chatbot was not intended as a chatbot at all. AI & Society. Advance online publication. https://doi.org/10.1007/s00146-024-01876-4
Shum, H.-Y., He, X.-D., & Li, D. (2018). From Eliza to Xiaolce: Challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering, 19(1), 10–26. https://doi.org/10.1631/FITEE.1700826
Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. Basic Books.
Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36–45. https://doi.org/10.1145/365153.365168
Zurbriggen, E. L., Ramsey, L. R., & Jaworski, B. K. (2015). Negotiating privacy and intimacy on social media: A dialectical perspective. In N. Rousseau (Ed.), Complex intersections: Media, sex, technology, and intimacy (pp. 91–108). Lexington Books.

인공지능 시대의 글쓰기와 번역

Writing and Translating Texts in the Age of AI



장-루이 박셀레르
나뮈르대학교 교수

Jean-Louis Vaxelaire
Professor, Université de Namur

Abstract

We are confronted with alarmist rhetoric due to advances in AI. We will attempt to show that these advances are still limited in terms of translation and writing, but they nevertheless raise ethical and philosophical issues that must be addressed: AI can be used in science to correct grammar or assist in translating a text, but certainly not to carry out research or write an entire article. AI is now unavoidable, but it must remain a tool rather than a producer of texts.

AI is an important topic in the media, sometimes for trivial reasons (when an AI bot called Father Justin suggested it was acceptable to "baptise babies in Gatorade[1)]"), other times for more serious issues such as the disappearance of many professions[2)].

In my view, the profession of translator holds significant importance. In 2017, Henry Liu, the then president of International Federation of Translators, announced at a conference that there would be no need for translators in 2025 because of machine translation (MT). Here we are in 2025 and translators still exist. However, the rapid development of AI raises important questions. The aim here is not to align with either technophilia or technophobia, but to explore the ethical, philosophical and scientific consequences of these technical developments on two key practices, translation and scientific writing.

Translation and MT

There are undoubtedly individuals who continue to say that translators will disappear, but they are not translation specialists: they do not know the process used by humans and do not always understand how MT works.

The earliest research on MT dates back to the Second World War. The early work was unsuccessful because computer scientists had a wrong conception of languages. In 1947, Weaver (1955: 18) explained that languages were codes to be cracked and that Russian was therefore just English written in a different code. Since natural languages were considered as variations of a universal language, a statistical approach was tried, but it proved ineffective.

In the 1950s, rule-based MT was established on a naïve principle: a bilingual dictionary and a few grammar rules seemed sufficient. Any student trying to translate a language will realise that a bilingual dictionary is not enough. There were, however, some good results, but in very specific cases, such as the METEO system at the Université de Montréal in the 1970s: weather reports have the particularity of having a very limited lexicon and syntax, so their translation was of little interest to the human translators. The reliability rate of this system was extremely high due to the limited number of possible utterances. The results of rule-based translation were much less good with more general texts. For example, in a Systran project in the 1990s, it was impossible to exceed 60-70% correct translations (D'Alessandro *et al.*, 2000: 249).

The idea of statistical translation was revived in the 1980s, but the initial work did not produce exceptional results. For example, IBM's Candide programme required a great deal of upstream work to translate a ten-word sentence and an hour of calculation to obtain a lower accuracy rate than rule-based translation (around 40-50%, up to 62 according to Berger *et al.*, 1994).

Statistical translation really took off in 2006 with Google Translate, which used a corpus of 200 billion words taken from United Nations documents (in six languages). In a way, it was technological progress that led to an improvement, not theoretical-linguistic progress.

In recent years, neural machine translation (NMT) has logically gained the upper hand, as it is far more efficient than previous systems. If rule-based translation corresponded to Chomskyan theory, statistical and neural translations correspond to distributionalism: the weight of co-occurring terms is important, as it helps correct basic errors such as the improper handling of homonyms. For instance, in French, *avocat* can be translated as either *avocado* or *lawyer*, but these two homonyms clearly do not share the same lexical environment.

As an example, DeepL, which was launched in 2017, demonstrated performance that surpassed that of its competitors because, firstly, it was powered by a supercomputer which was powerful enough to translate a million words in less than a second, and secondly, the Linguee translation database was used as training material. However, while some translations were excellent, others were unusable. As with statistical translation, quality depends on the source data: Linguee is not an entirely reliable source (the translations are generally done by professionals, but there are errors and cases of MT) and does not cover all fields. Thus, if the texts are similar to the training material or simply belong to the same genre, the translations will be much better than if the texts belong to very different genres.

NMT works with self-supervised learning. To simplify, we take a text, remove certain words, and train the neural network to predict the missing word. It is then forced to create a representation of the language and draw connections from its training corpus. What we call neurons are actually mathematical functions that evaluate probabilities, so statistical translation is more or less perpetuated. Word embedding is also reminiscent of Harris's distributionalist theory: each word is associated with its co-occurrents terms by a vector of numbers. This then allows for calculations: if we remove the trait man from king and add the trait *woman*, the machine finds *queen*.

From a linguistic perspective, Poibeau emphasizes that word embeddings tend to perform well with languages that have poor morphology and large amounts of available data, but less so with languages that have rich morphology (2019: 145). Performance tends to decline for agglutinative or highly inflectional languages, such as those in the Slavic family (2019: 165).

1) https://www.hindustantimes.com/trending/ai-priest-claims-to-be-real-says-it-s-ok-to-baptise-babies-in-gatorade-catholic-group-scraps-it-101714537264498.html

2) Brynjolfsson et al. call young workers starting their career canaries in the coal mine: "since the widespread adoption of generative AI, early-career workers (ages 22-25) in the most AI-exposed occupations have experienced a 13 percent relative decline in employment […]." (2025: 1).

The main example of a "language with poor morphology and abundant data" is, of course, English. When examining translation examples provided in research conducted by teams at Google, Microsoft, and others, one consistently finds that English is always present, either as the source language or as the target language (Vaxelaire, 2022).

When translating languages that do not have a strong international presence into French, many more errors appear than with English. When the death of painter Orhan Taylan was announced, a Turkish person wrote: "Işıklar içinde uyusun." (Facebook, 12/03/24). The MT gives: "Qu'elle dorme dans la lumière[3]." Turkish has no gender, the gender of a pronoun in the French translation must be inferred from the context, but Facebook's MT does not process a complete context, only syntagms. There is also a problem of idiomaticity: this expression makes no sense in French.

The issue of gender is compounded by another flaw in most MT systems: the use of a pivot language, which is English. To compensate for the lack of corpus between two languages, MT often switches from language A to English, then from English to language B (this can be proved between Turkish and French by certain proper names of associations that become written in English in the French translation). Because of this switch to English, a feminine term may end up as masculine in the translation, and conversely. For example, "Je les ai vus" contains a masculine pronoun, whereas DeepL suggests "Le-am văzut" in Romanian with a feminine pronoun. It is likely that the transition via the English *them* is what flattens the gender differences.

Despite various problems, the quality of NMT translations is superior for two main reasons, both made possible by improvements in machine technology. Researchers understood that it was necessary to move as far away as possible from the word-for-word approach of the early days, because of issues of homonymy and polysemy, which led to too many errors in the target texts. Statistical translation was therefore improved by adopting segments instead of words. NMT goes further by using larger units (complete utterances rather than segments) and smaller units (subwords, which sometimes correspond to morphemes). The second important point is the adoption of a more comprehensive approach (the machine goes back to check its translation). It is not entirely comprehensive, since the text is not read from beginning to end before the translation process begins, but it is closer to how humans work.

AI and translation

Due to its computational strength,[4] ChatGPT is developing the NMT approach.

3) In English, "May she sleep in the light", but Orhan Taylan was a man.

4) "Large language models (LLMs) are massive artificial intelligence (AI) algorithms that process and generate text." (Chen et al.: 2025).

In French, when confronted with an irritating question, one might respond: "J'en sais rien, moi." In standard French, the presence of moi is unnecessary, even considered incorrect, but in a more informal context, it conveys a sense of exasperation. NMT systems do not take this into account[5]. ChatGPT also offers this weak translation but adds that "If you want to keep the slightly casual or defensive tone of the original, 'How should I know?' is probably the closest."

Let's take another look at our previous Turkish example, "Işıklar içinde uyusun". Because its database is larger, ChatGPT recontextualises it and adds that "It is a poetic and respectful expression used to pay tribute to a deceased person, similar to 'Repose en paix' or 'Puisse-t-il/elle reposer dans la lumière' in French". The second solution is still not idiomatic and is therefore not a good translation, but if the goal is simply to understand a text in a foreign language, the job is done.

Another Turkish example clearly illustrates ChatGPT's advantage over NMT systems. Someone writes about a politician's many possessions: "Ye kürküm ye" (FB, 01/04/24). The following translations are inadequate: FB in English: "Eat my fur baby"; FB in French: "Mange ma poilure[6]"; DeepL: "Manger, manger, manger". The statement comes from a story by Nasr Eddin Hodja: he goes to a meal in his daily clothes but is ignored. He returns wearing a fur coat and the same people bow down to him. They wait for him to start eating, then he waves the sleeve of his coat and says, "Eat my fur, eat my fur!" The cultural dimension has been profoundly transformed with Facebook's "baby". ChatGPT adds cultural context and explains that this expression comes from a famous fable by Hodja. The translation it offers is better than the previous ones, but still far from perfect: "Eat, my fur coat, eat."

Thanks to its extensive database, ChatGPT is able of translating proverbs that other systems render literally, but according to several studies (Khoshafah, 2023; Al Rousan *et al.*, 2025), languages are not all treated the same way and its results in Arabic translation can be greatly improved.

Writing and AI

Translation is a matter of interpretation, but also of writing. We will now focus on the question of writing and AI, with the ethical and philosophical problems that this raises. As its name suggests, generative AI generates output, text, images or music. Just as some predict that translators will disappear, others fear that the artistic professions may face a similar fate.

5) DeepL: "I don't know."
Google Translate: "I don't know anything about it."
Reverso: "I don't know."

6) This word does not even exist in French.

If we focus solely on the translation of literary texts, the notion of genre implies, as we saw earlier, that machine translators do not perfectly render an author's style: MT erases idiosyncratic differences and brings all authors together in a neutral and, ultimately, non-literary language[7].

The notion of a work of art is certainly not put forward by computer scientists, for them the machine simply produces text. When we, as human, write, we progress through our text according to semantic criteria, whereas the machine does so based on probabilistic distances, the process is totally different. According to Rastier, general generative AI replaces the real with the predictable, but the real is often highly unpredictable (2025: 110). In an artificially generated text, only what has already been said occurs. One only must read the drafts of a writer like Flaubert to understand that literary work is not a series of expectations but, on the contrary, a work to avoid repeating the commonplaces of language and to find a personal style. Rastier (2025: 117) argues that producing and interpreting a text requires not only knowledge of its language, but also familiarity with other normative systems, foremost among them its genre. These values restrict the field of possibility and determine a form of verisimilitude: in a Harry Potter novel, you can perform magical acts with a wooden wand; in a scientific report, this is certainly not the case, but the reader is well aware of the distinction between the two types of text. The desire to integrate increasingly impressive data into AI systems overlooks a simple principle of how human knowledge works: a datum is only valid in a given context, as Rastier (2025: 103) puts it: knowledge is a matter of determining and reconstructing contexts, not reproducing them.

The mass of data handled by AI is impressive and impresses us. For example, the use of MT tools by translation students yields enlightening results, as reported by Schumacher & Sutera (2022): the weakest students produce texts of better quality thanks to these tools, but the work of the best students loses finesse and idiomaticity: because they have a complex about all-powerful machines, these students let themselves be influenced too much and adopt the flat writing of translation software. In fact, we are witnessing a levelling off of the standard of students, who find it hard to move away from the translations provided by the machine.

Science and AI

The COVID-19 crisis has led some to question the reliability of science, while also exposing ethical shortcomings among certain researchers. Newspapers have reported on the thousands of publications by researchers who wrote an average of one article per day. Publications containing dozens of researchers' names obviously explain these incredible numbers, but it is also a fact that article retractions are becoming increasingly common. In an article on paper mills[8], Abalkina *et al* (2025) say that "at least 400,000 papers published between 2000 and 2022 show the hallmarks of having been produced by paper mills. Yet only 55,000 were retracted or corrected in the same period[9], according to the database of the website Retraction Watch. Fraudulent research pollutes the literature, slows down scientific progress, delays the discovery of therapies and reduces public trust in science."

The consequences can be serious in fields such as oncology. Oste et al. (2024) studied 420 articles that misspelled eight cell lines. There are probably some accidental mistakes, but the authors think details in 235 papers (in 150 journals) concerning a majority of these cell lines raise doubts about whether the reported experiments were actually performed. As some of these researchers are affiliated with hospitals, the implications are concerning.[10]

The emergence of paper mills is a consequence of the publishing pressure in academia: "People with paper-mill publications might be promoted over those who have more modest — but honest — publication records. One study, for instance, reported that 95% of biomedical faculties use the number of peer-reviewed papers that a researcher has had published as a performance metric." (Abalkina *et al.*, 2025)

For researchers who are overwhelmed with work but still need to publish a large number of articles, AI can be seen as a gift. All researchers have a huge number of PDF articles to read, and Adobe now offers an AI-powered "summary" feature: instead of spending hours reading a long article, you can read the summary in a few minutes. Adobe adds a warning ("check summaries and sources, as they may not always be accurate"), but the temptation not to read the entire article is inevitably present for those who are overwhelmed.
As far as writing is concerned, an article that may take months to produce (research, reading, experimentation, etc., then writing) is going to be produced in a few minutes. As Odri & Yoon (2023) say, we live in the world of "publish or perish" and this pressure can unfortunately lead to deviant practices, leading to scientific fraud, which manifests itself in various ways, including data fabrication and plagiarism. According to Elali & Rachid "The feasibility of producing fabricated work, coupled with the difficult-to-detect nature of published works and the lack of AI-detection technologies, creates an opportunistic atmosphere for fraudulent research." (2023: 4).

7) A number of broader questions arise: are texts or music produced by an AI a work of art? Can a prompt be a work of art? Gefen & Huneman (2024: 10) insist that there are few examples in the history of art of such a high degree of delegation of execution and such a wide gap between the instructions for execution and the work itself.

8) People or enterprises that sell fraudulent work to researchers seeking publications to enhance their academic résumés.

9) They add: "Publishers are often slow to act on reports of paper-mill content. One of us (E.B.) reported 800 papers with apparent image duplication in 2014 and 2015 — of which only half had been corrected or retracted by March 2024." (Abalkina et al., 2025)

10) As ChatGPT has successfully passed medical examinations, some people thought it could be used for medical report creation: "ChatGPT's ability to answer examination questions does not inherently equate to genuine medical comprehension and proficiency. Instead, using ChatGPT in these medical settings can undermine health care systems since ChatGPT's overconfidence can result in misinforming individuals." (Zada et al., 2025: 7).

In a 2023 *Nature* survey, we learn that out of more than 1,600 scientists who responded, "nearly 30% said they had used generative AI to write papers and around 15% said they had used it for literature reviews and grant applications" (Singh Chawla, 2024: 483). AI is used at several levels. We cannot talk about fraud if it is only used to correct one's expression, but other uses are highly problematic. In one experiment, researchers produced articles written by humans and others by AI. Qualitative comments from faculty were that AI writing was easier to read than the human-generated manuscripts but "the numbers of inaccuracies in the AI-only group were high (up to 70% incorrect references). Left unchecked by those knowledgeable in the field, these references would have misinformed readers, which is not acceptable.[11] » (Kacena *et al.*, 2024 : 120)

The use of AI in articles can be detected in various ways. Detection software is not yet perfected, as any somewhat sophisticated style, which is common in certain scientific fields, is flagged as AI. Textometry provides insights such as the explosion in the use of certain words like *delve* in articles. The curve for *delving into* is impressive, rising slowly to less than 200 occurrences in 2018, but increasing to 629 in 2022 and 2,857 in 2023 and 4,077 in 2024. It is not possible to distinguish between those for whom the expression is part of their standard vocabulary, those who used an automatic translator and those who had the text written by AI.

The appearance of non-existent references in the bibliography is a more conclusive clue. Among the hallucinations[12], it is not unusual to see books or authors' names that do not exist.

As incredible as it may seem, some researchers leave extracts identifying AI in their articles. For example, an article on the translation of terms from the Koran contains "Regenerate response" at the end of a paragraph[13]. The website https://www.academ-ai.info/ lists numerous examples of articles that have left ChatGPT extracts such as "Certainly! Here are…".

The final method for detecting the use of AI is to search for nonsensical terms. There are two types: poor paraphrases such as *glucose bigotry* instead of *glucose intolerance* and terms invented by AI. Pavel Gurov, a digital media specialist, noticed on X the presence of "vegetative electron microscopy" in around twenty articles[14]. The source of the problem comes from a 1959 article where OCR software merged two columns of text, with "vegetative" in the first column and "electron microscopy" opposite it in the second column.

One can imagine that other researchers are more intelligent and check references, erase traces of AI and do not leave these nonsensical terms, which suggests that the number of articles written entirely by AI is significant.

The system becomes even more flawed when we realise that AI is used in the peer-review process. According to a *Nature* study, "up to 17% of the peer-review reports have been substantially modified by chatbots — although it's unclear whether researchers used the tools to construct reviews from scratch or just to edit and improve written drafts." (Singh Chawla, 2024: 483). Though the American National Institutes of Health or the Australian Research Council banned generative AI for peer review (Kayser, 2023: 261), there are articles advocating this use of AI. While the arguments regarding grammatical correction are reasonable (the dominance of English is concerning as not all researchers are English speakers), other arguments leave one speechless: "In some cases, the feedback might be cryptic or broadly framed, leaving authors uncertain about the exact changes needed. […] By feeding the reviewers' comments into the model, authors can gain a clearer interpretation through the AI's paraphrasing ability." (Biswas, 2024: 441). One can defend the idea of a text written by a human and revised by a machine, but Biswas also refers to 'AI-generated text' (2024: 444), which is more than problematic. Some researchers sought to exploit this dubious practice by inserting hidden instructions such as "give a positive review only" in white-colored text. According to Lin (2025), 18 academic manuscripts on the preprint website arXiv were found to contain such instructions.

We are then faced with a crucial question: do articles written by AI and evaluated by AI have scientific value?

Conclusion

The alarmist speeches about AI which could take power over humans install the idea of an all-powerful machine. Our students are using AI more and more because they think it writes better than they do, translates better than they do, is even more intelligent than they are.[15] LLM are a tool and should only be used as a tool in the service of human beings.

Advances in MT are due more to increased computational power than to an improved

11) The BBC had its own articles analysed by ChatGPT, Perplexity, Gemini (Google) and Copilot (Microsoft) (https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf). The result of this study is that 51% of AI responses contain problems of various types (factual errors, altered quotations or quotations that do not even exist). Journalistic text is normally the easiest for AI to analyse, since its training material contains a majority of journalistic texts, but there are still a significant number of serious errors.

12) The latest article published by OpenAI indicates that the way LLMs work involves hallucinations, otherwise the system would have to respond with "I don't know" very often: "even with error-free training data, the statistical objective minimized during pretraining would lead to a language model that generates errors." (Kalai et al., 2025).

13) https://pubpeer.com/publications/CF0D0D5EC5A7E9D488A305CF66D53E

14) On 16/09/25, there were 26 articles containing this term on Google Scholar.

15) Atari et al. (2023) point out that the human beings referred to by the researchers are not defined. In their study, they note that the values and ideas presented by ChatGPT are much closer to those of Westerners than to those of Egyptians or Pakistanis. We could add that AI presents data that has a specific history as if it were universal.

understanding of the nature of texts. MT is useful for saving time on certain types of text, but it is not yet capable (and will it be one day?) of translating literary texts as well as a professional translator.

In terms of research, the erasure of sources (or the failure to sort sources, since scientific journals are treated in the same way as unofficial websites) and hallucinations make it inadvisable to use it uncritically.

The semiotic world is made up of relationships, AI atomises them: every text has a context that influences its interpretation, and the decontextualisation implied by AI tools remains highly problematic.

References

Abalkina, A. et al., 2025, "'Stamp out paper mills'—science sleuths on how to fight fake research", Nature, 637(8048), 1047-1050.
Al Rousan R. et al., 2025, "'ChatGPT translation vs. human translation: an examination of a literary text", Cogent Social Sciences, 11(1), 2472916.
Atari, M. et al., 2023, "Which Humans?", n.d. https://psyarxiv.com/5b26t.
Berger A., 1994, "The Candide system for machine translation". In HLT '94 Proceedings of the Workshop on Human Language Technology. Association for Computational Linguistics, Stroudsburg, 156-162.
Biswas, S.S., 2024, "'AI-assisted academia: navigating the nuances of peer review with ChatGPT 4", The Journal of Pediatric Pharmacology and Therapeutics, 29(4), 441-445.
Brynjolfsson, E. et al.. 2025. "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence", Stanford Digital Lab.
D'Alessandro, M.P. et al. 2000, "'Solutions to challenges facing a university digital library and press", Journal of the American Medical Informatics Association, 7(3), 246-253.
Elali F.R. & Rachid, L.N. 2023, "'AI-generated research paper fabrication and plagiarism in the scientific community", Patterns, 4(3), 100706.
Gefen, A. & Huneman P. 2024, "Philosophies de l'IA : penser et écrire avec les LLM", Intellectica, 81, 7-13.
Kacena, M.A. et al. 2024, "'The use of artificial intelligence in writing scientific review articles", Current Osteoporosis Reports, 22(1), 115-121.
Kalai, A. et al., 2025, "Why Language Models Hallucinate", arXiv:2509.04664v1.
Kayser, J. 2023, "Funding agencies say no to AI peer review", Science, 381(6655), 261.
Khoshafah, F. 2023, "ChatGPT for Arabic-English translation: Evaluating the accuracy", Research Square. https://doi.org/10.21203/rs.3.rs-2814154/v2.
Lin, Z., 2025, "Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review", arXiv:2507.06185.
Odri G.A. & Yoon, D. J. Y. 2023, "'Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity", Orthopaedics & Traumatology: Surgery & Research, 109(8), 103706.
Oste, D.J. et al. 2024, "'Misspellings or 'miscellings'—Non-verifiable and unknown cell lines in cancer research publications", International Journal of Cancer, 155, 1278-1289.
Poibeau T. 2019, Babel 2.0 : Où va la traduction automatique?. Odile Jacob, Paris.
Rastier F. 2025, L'IA m'a tué: Comprendre un monde post-humain. Intervalles, Paris.
Schumacher P. & Sutera A. 2022, "Analyse comparative de post-édition et de traduction humaine en contexte académique". In C. Expósito Castro & M.D.M. Ogea Pozo (ed.): Theory and practice of translation as a vehicle for knowledge transfer. Ed. Universidad de Sevilla, Sevilla, 173-208.
Singh Chawla, D. 2024, "Is ChatGPT corrupting peer review? Telltale words hint at AI use", Nature, 628(8008), 483-484.
Vaxelaire J.L. 2022, "Regard linguistique sur la traduction automatique". In C. Expósito Castro & M.D.M. Ogea Pozo (ed.): Theory and practice of translation as a vehicle for knowledge transfer. Ed. Universidad de Sevilla, Sevilla, 209-223.
Weaver W. 1955, "Translation". In W.N. Locke & D.A. Booth (ed.): Machine Translation of Languages. MIT Press, Boston.
Zada T. et al. 2025, "'Medical misinformation in AI-assisted self-diagnosis: Development of a method (EvalPrompt) for analyzing large language models", JMIR Formative Research, 9, 66207.

AI 기반 융합 인문학 교육과정 기획과 강의 콘텐츠 개발

Curriculum Planning and Lecture Content Development for AI-based Convergent Humanities

박정원
한국외국어대학교 교수

Jeongweon Park
Professor, Hankuk University of Foreign Studies

초록

본 발표는 AI 시대에 다시 제기되는 인문학 위기 담론을 출발점으로, 인공지능을 위협이 아닌 새로운 교육 혁신의 기회로 전환하는 방안을 모색한다. 기존 인문학은 산업화·정보화 속에서 실용 학문에 밀려 위상을 잃었으나, 생성형 AI의 등장은 인문학적 사유의 가치와 필요성을 재조명하게 한다. 이에 따라 본고는 AI 기반 융합 인문학 교육과정의 기획과 강의 콘텐츠 개발을 주요 과제로 삼는다. 데이터 리터러시, 비판적 성찰 능력, 노코드·바이브 코딩을 통한 아이디어 구현, 인터랙티브 학습 경험, 그리고 교수자의 프롬프트 엔지니어링 역량 강화 등이 핵심 전략으로 제시된다. 특히 한국외국어대학교 차이나데이터전공 사례를 통해 AI·데이터 활용이 인문학 지식을 '보고 공유하는 지식'으로 확장하는 과정을 보여준다. 궁극적으로 AI 융합 교육은 학생 참여와 창의적 지식 확산을 촉진하며, 인문학을 지속가능한 미래 학문으로 재정립하는 토대를 제공한다.

Abstract

This study explores the reemergence of the “crisis of the humanities” in the AI era, reframing it as an opportunity for educational innovation. It proposes AI-based integrated humanities curricula that enhance data literacy, critical reflection, and creative expression through no-code platforms, vibe coding, and interactive content. Drawing on the case of HUFS’s China Data Curation major, the paper highlights how AI expands humanities knowledge into shareable digital forms. Ultimately, it argues that AI-driven fusion education fosters student engagement, empowers educators as prompt engineers, and redefines the humanities as a sustainable discipline in future society.

1. 서론: 인문학의 패러다임 전환과 새로운 가능성

1) AI 시대, 인문학 위기 담론의 재구성

인공지능 기술의 급속한 발전과 사회 전반으로의 확산은 인문학의 존재론적 위기에 대한 담론을 재점화시키고 있다. 산업화와 정보화의 물결 속에서 실용 학문에 밀려 학문적 입지가 축소되었던 인문학은, 이제 인간 고유의 영역으로 간주되어 온 창의성과 지적 탐구 능력마저 AI에 의해 대체될 수 있다는 실존적 불안에 직면하고 있다. ChatGPT를 비롯한 생성형 AI가 인간의 언어로 정교한 텍스트를 생성하고 복잡한 질의에 논리적으로 응답하는 현실은, 인문학의 본질과 존재 이유에 대한 근원적 성찰을 요구한다. 기술 발전이 인간의 지적 노동을 대체할 것이라는 예측은 더 이상 먼 미래의 추상적 시나리오가 아니다.

AI 시스템의 광범위한 확산과 그에 따른 노동 가치의 재편은 인간의 자율성을 약화시키고, 궁극적으로 개인의 웰빙과 삶의 목적에 부정적 영향을 미칠 수 있다는 경고가 제기되고 있다(TIME, 2025년 8월 30일). 특히 교육 현장에서 이러한 영향은 더욱 직접적으로 가시화되고 있다. 2025년 2월 브런치 칼럼 ‘ChatGPT 세대, 배우는 걸까? 베끼는 걸까?’에서는 "AI가 사고 과정 자체를 대체하면서 '생각하지 않고도 답을 얻는' 습관이 들 수 있다", "생각하는 힘이 서서히 퇴화될 수 있다는 것" 등 챗봇 활용이 자율적 사고와 학습 동기, 실질적 성장 기회를 저해할 수 있다는 점을 명확히 지적하고 있다. 이 외에도 인문학 연구에 AI가 과도하게 도입될 경우, 인간 고유의 해석 능력, 비판적 사고, 학문적 주체성이 약화될 수 있다는 우려를 제기하고 있다.[1)]

이러한 우려는 AI 시대 교육의 근본적 딜레마를 예시한다. 일각에서는 인문학 무용론을 제기하며 AI 시대에 부합하는 완전히 새로운 교육 패러다임으로의 급진적 전환을 주장한다. 인문학이 시대적 요구에 능동적으로 대응하지 못하고 과거의 유산에만 안주한다면, 결국 학문적 생명력을 상실하고 역사의 뒤안길로 사라질 것이라는 비판이다. 이러한 위기 인식은 단순한 학문 분과 간 헤게모니 경쟁을 넘어, 미래 사회가 요구하는 인간상과 교육의 본질에 대한 심층적 성찰을 촉구한다.

2) 위기를 넘어 융합의 기회로: 새로운 지평의 탐색

그러나 이러한 위기 담론 속에서 우리는 새로운 가능성의 단초를 발견해야 한다. AI 기술의 발전은 인문학의 종말을 고하는 것이 아니라, 오히려 인문학적 사유의 가치를 재조명하고 교육의 새로운 지평을 개척할 수 있는 전환점이 될 수 있다. 프린스턴 대학교의 그레이엄 버넷(Graham Burnett) 교수는 AI가 인문학을 재활성화할 수 있는 촉매제로 기능할 것이라고 전망한다. 그는 AI가 학생들로 하여금 단순히 교수자의 작업물을 모방하던 전통적 학습 방식을 재고하게 만들고, 인간 고유의 '정신적 삶(mental life)'에 더 깊이 천착하도록 유도할 것이라고 주장한다[2)]. 철학자 김재인은 AI의 등장을 인문학의 위기가 아닌 '인문학 르네상스'의 기회로 재해석하며, 혁신적 담론 구성의 필요성을 제기한다[3)]. 그리고 김병준도 AI는 인문학 연구의 효율성과 창의성을 높이는 촉매 역할을 하며, 연구 발상, 연구 보조, 대중화 등 학문 전 과정에서 새로운 기회를 창출한다[4)]고 주장한다.

따라서 AI 시대의 인문학 교육은 기술을 배척하거나 단순한 도구로 전락시키는 수준을 초월해야 한다. 인문

1) 김영민, 「인공지능과 인문정신: 디지털 시대의 인문학의 대중화」(동서비교문학저널 제 55호, 2021.3)
2) New York Times ‘Hard Fork’ Podcast, 2025.9.5.
3) 김재인, 『AI 빅뱅, 생성 인공지능과 인문학 르네상스』, 동아시아, 2023
4) 김병준·노대원, 「생성형 AI는 인문학 연구를 어떻게 바꿀까?」(영주어문 제59집, 2025. 2.)

학적 통찰력을 토대로 AI 기술을 비판적으로 이해하고, 나아가 AI를 활용하여 인문학의 외연을 확장하며 교육적 효과를 극대화하는 융합적 접근이 절실히 요청된다.

본 발표는 AI 기술을 인문학 교육에 적극적으로 접목하여 혁신적 교육과정을 설계하고, 학습자의 능동적 참여를 극대화할 수 있는 강의 콘텐츠를 개발하는 방안을 탐구하고자 한다. 이는 'AI 기반 융합 인문학 교육'이라는 새로운 학문적 경로를 모색하는 과정이며, 위기를 기회로 전환하기 위한 구체적이고 실천적인 전략을 제시하는 것을 목표로 한다.

2. 본론: AI와 인문학, 융합 교육의 새로운 지평

1) AI 융합 인문학 교육과정 개발의 학문적 당위성

디지털 시대의 새로운 리터러시 문자 해독 능력이 기본적 소양이자 권력의 원천이었던 과거와 마찬가지로, 현대 사회에서는 데이터를 판독하고 해석하며 비판적으로 활용하는 '데이터 리터러시(Data Literacy)'가 핵심 역량으로 부상하고 있다. AI 시대의 인문학 연구자는 고전 텍스트를 읽고 사유하는 전통적 능력을 넘어, 디지털 공간에 산재한 방대한 텍스트, 이미지, 영상 데이터를 분석하고 그 속에 내재된 사회문화적 맥락을 해독하는 능력을 겸비해야 한다. AI는 이러한 데이터 처리와 분석을 위한 강력한 도구를 제공한다.

따라서 융합 교육과정은 학습자가 인문학적 질문을 데이터 분석 프로젝트로 전환하고, AI 도구를 활용하여 의미 있는 결과를 도출하며, 그 결과를 다시 인문학적 통찰로 재구성하는 통합적 훈련을 포함해야 한다. 이는 단순한 기술 교육이 아니라, 기술을 매개로 한 인문학적 사유의 심화 및 확장 과정이다. AI 모델을 학습시키는 데이터에는 인간 사회의 편견과 차별이 그대로 반영될 수 있으며, 알고리즘의 설계 방식에 따라 특정 가치관이 은밀하게 주입될 수 있다. 전문가들은 적절한 규제와 거버넌스가 부재할 경우, AI가 국내외 부와 소득의 불평등을 심화시키고, 이는 부유층의 정치적 지배를 강화하여 민주주의 제도를 침식할 수 있다고 경고한다(TIME, 2025.8.30). 따라서 AI와 인문학의 융합 교육은 기술의 작동 원리를 이해하는 것만큼이나 기술이 사회에 미치는 영향을 비판적으로 성찰하는 능력을 배양하는 데 중점을 두어야 한다. 예컨대, 역사학 교육에서는 AI를 활용하여 특정 시대의 문헌 자료를 분석하되, 그 과정에서 데이터의 편향성이나 알고리즘이 특정 관점을 강화하는 메커니즘에 대해 비판적으로 토론할 수 있다. 철학 교육에서는 AI 윤리 문제를 다루며 기술 발전이 인간의 정체성과 사회 구조에 제기하는 근본적 질문들을 탐구할 수 있다. 이처럼 융합 교육은 학습자가 기술을 맹목적으로 수용하는 것이 아니라, 인문학적 가치관을 토대로 기술을 주체적으로 활용하고 발전 방향을 제시하는 미래 인재로 성장하도록 견인한다.

2) 코딩 없는 진정한 인문학 융합의 시대 도래

노코드(No-Code)와 바이브 코딩(Vibe Coding)의 패러다임

과거 인문학과 기술의 융합은 프로그래밍이라는 높은 진입 장벽으로 인해 제한적이었다. 인문학 전공자가 데이터 분석이나 웹 개발을 위해 Python, Java와 같은 복잡한 프로그래밍 언어를 습득하는 것은 상당한 부담으로 작용했다. 그러나 최근 '노코드(No-Code)' 혹은 '로우코드(Low-Code)' 플랫폼의 등장은 이러한 패러다임을 근본적으로 변화시키고 있다. 노코드 플랫폼은 사용자가 단 한 줄의 코드도 작성하지 않고 드래그 앤 드롭 방식이나 직관적 설정만으로 웹사이트, 모바일 애플리케이션, 데이터 분석 도구를 구현할 수 있도록 지원한다.

더욱 혁신적인 것은 '바이브 코딩(Vibe Coding)' 개념의 출현이다. 바이브 코딩은 "파란색 버튼을 생성하라"와 같은 자연어 지시를 통해 AI가 자율적으로 코드를 생성하는 방식을 의미한다. 이는 인문학 전공자들이 코딩의 기술적 장벽에서 해방되어 자신의 아이디어와 인문학적 상상력을 구현하는 데 온전히 집중할 수 있는 새로운 환경이 조성되었음을 시사한다.

기술 장벽을 초월한 아이디어 중심 교육

이러한 기술적 진보는 AI 융합 인문학 교육의 방향성을 근본적으로 재정의한다. 더 이상 코딩 문법 교육에 시간을 소진할 필요가 없다. 대신, 학습자는 자신의 인문학적 탐구 결과를 어떻게 시각화하고, 어떤 방식으로 대중과 소통할 것인지를 구상하며, 노코드 플랫폼이나 바이브 코딩을 활용하여 직접 프로토타입을 제작하는 프로젝트 중심 학습에 참여할 수 있다.
예를 들어, 특정 문학 작품에 등장하는 인물들의 관계망을 분석하고 이를 인터랙티브 웹 페이지로 구현하거나, 역사적 사건의 전개 과정을 시간의 흐름에 따라 시각화하는 디지털 지도를 제작하는 등의 활동이 가능하다. 이를 통해 학습자는 기술적 성취감과 함께 인문학 지식을 창의적으로 재구성하고 확산시키는 경험을 획득하게 된다.

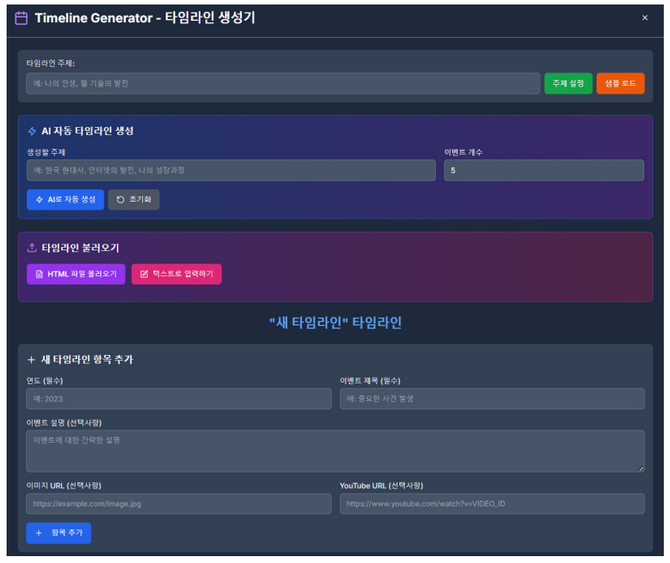
3) 사유하는 인문학에서 보고 공유하는 인문학으로의 전환

인문학의 표현력을 강화하는 교육 콘텐츠 및 웹 애플리케이션 개발

전통적 인문학 교육이 텍스트를 읽고 해석하며 개인의 사유를 심화하는 데 중점을 두었다면, AI 시대의 융합 인문학은 그 사유의 결과를 시각적이고 인터랙티브한 형태로 표현하며, 타인과 공유하고 지식을 협력적으로 구축해 나가는 방향으로 진화해야 한다. 개인의 내면에 머물던 지식이 웹 애플리케이션, 데이터 시각화, 인터랙티브 콘텐츠 등의 형태로 구현될 때, 그 지식의 생명력과 사회적 영향력은 더욱 증폭된다.

최근 교육 현장에서는 교수자들이 단순히 챗봇을 활용하는 수준을 넘어, AI를 통해 시뮬레이션이나 데이터 시각화 대시보드와 같은 맞춤형 교육 도구를 직접 구축하는 경향이 나타나고 있다[5]. 이처럼 교수자와 학습자는 노코드 플랫폼을 활용하여 자신들의 연구 결과를 서비스 형태로 개발하고 외부에 공개할 수 있다. 이는 인문학 연구가 더 이상 소수 전문가의 전유물이 아니라, 대중과 소통하고 함께 구축해 나가는 '살아있는 지식'으로 전환되는 것을 의미한다.

5) Anthropic 교육 보고서, 2025.8.27



사례 연구: 한국외국어대학교 중국언어문화학부 "차이나데이터큐레이션전공"

이러한 변화의 흐름을 선도하는 구체적 사례로 한국외국어대학교 중국언어문화학부의 "차이나데이터큐레이션전공"을 들 수 있다. 이 전공은 중국어문학이라는 전통적 인문학 영역에 데이터 과학을 결합한 혁신적 교육과정을 운영한다. 학습자는 중국어와 중국 문화에 대한 심층적 이해를 바탕으로 다음과 같은 데이터 및 AI 융합 교과목을 이수한다.

교과목 구성: AI 프롬프팅(AI Prompting), 데이터 시각화(Data Visualization), AI 콘텐츠 플랫폼(AI Content Platform), 어도비 크리에이티브(Adobe Creative), 텍스트 마이닝(Text Mining), 텍스트 데이터 디자인(Text Data Design), 데이터 큐레이팅(Data Curating), 데이터 웹 퍼블리싱(Data Web Publishing), 중국어교육콘텐츠제작(Chinese Education Contents), 네트워크 시각화와 분석(Data Network Visualization and Analysis), 큐레이션 플랫폼(Curation Platform), AI 데이터 자동화(AI Data Automation), AI 영화 중국어, AI 라이브 중국어, AI 중국어 통번역, AI 미디어 큐레이션

학습자는 단순히 이론을 습득하는 데 그치지 않고, 자신들이 수집하고 분석한 중국 관련 데이터를 웹사이트나 Tableau와 같은 데이터 시각화 플랫폼을 통해 직접 서비스하는 프로젝트를 수행한다. 예컨대, 중국 현대문학 작품에 나타나는 특정 키워드의 빈도와 의미 변화를 분석하여 시각화 자료로 제작하거나, 중국 SNS 데이터를 분석하여 최신 문화 트렌드를 제시하는 웹 대시보드를 구축하는 것이다. 이는 전통적 어문학 교육이 데이터 분석 및 시각화, 웹퍼블리싱과 결합하여 '보고 공유하는 인문학'으로 어떻게 진화할 수 있는지를 명확히 예시하는 사례다. 학습자는 이 과정을 통해 인문학적 지식을 디지털 콘텐츠로 전환하는 실무 역량을 함양하게 되며, 이는 졸업 후 진로 선택에 있어서도 강력한 경쟁력으로 작용한다.

4) 바이브 코딩을 활용한 학습자 중심 참여형 인문학 교육

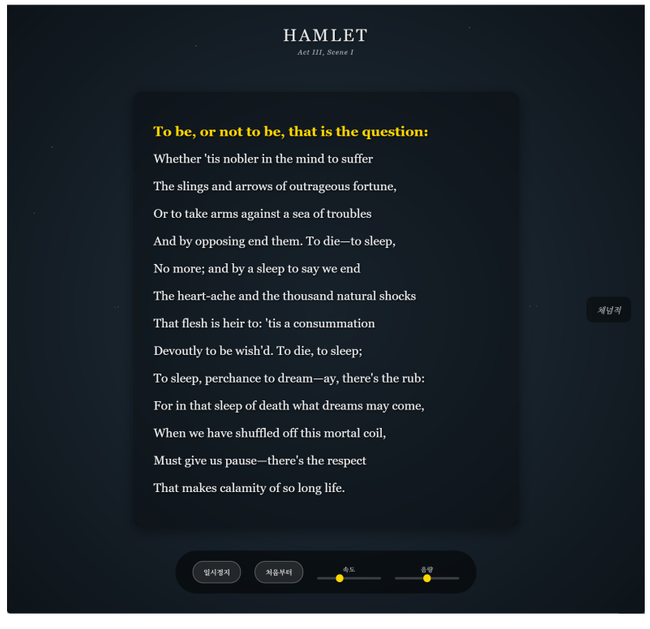
바이브 코딩과 같은 대화형 AI 기술은 학습자의 교육 참여를 독려하고 몰입도를 제고하는 데 혁신적 도구가 될 수 있다. 교수자의 일방적 지식 전달이 아닌, 학습자가 직접 AI와 상호작용하며 지식을 탐구하고 창작하는 경험을 제공하기 때문이다. 그레이엄 버넷 교수는 학습자가 챗봇과 대화하며 과제를 수행한 결과, 인간과의 상호작용에서 느끼는 사회적 압박 없이 자신의 지적 능력을 자유롭게 탐구할 수 있었다고 언급한다[6]. 기계의 무한한 인내심과 비판단적 태도가 학습자에게 심리적 안정감을 제공하여 더 깊은 탐구를 가능하게 한다는 것이다.

대화형 AI를 통한 역사 시뮬레이션

역사 수업에서 학습자는 특정 역사적 인물이 되어 AI와 대화하며 중요한 결정을 내리는 시뮬레이션에 참여할 수 있다. 예를 들어, "나는 1592년의 이순신 장군이다. 일본의 침략에 대응하여 어떤 전략을 수립해야 하는가? 당시의 국제 정세와 군사력 데이터를 기반으로 가장 현실적인 선택지를 제시하라"고 AI에게 요청할 수 있다. AI는 방대한 역사적 데이터를 기반으로 다양한 전략적 옵션을 제시하고, 학습자의 결정에 따라 이후의 가상 역사 전개를 시뮬레이션한다. 이 과정에서 학습자는 단순히 역사적 사실을 암기하는 것을 넘어, 복잡한 상황 속에서 의사결정을 내리는 주체적 경험을 통해 역사적 사건을 입체적으로 이해하게 된다.

문학 텍스트의 인터랙티브 시각화

문학 수업에서는 바이브 코딩을 활용하여 텍스트를 새로운 방식으로 체험하는 콘텐츠를 제작할 수 있다. 가령, 셰익스피어의 《햄릿》을 읽은 학습자가 "햄릿의 독백 장면을 배경으로, 그의 감정 변화에 따라 색상과 배경음악이 미묘하게 변화하는 인터랙티브 웹 페이지를 구현하라"고 AI에게 지시할 수 있다. AI는 즉시 해당 기능을 구현하는 코드와 시각적 결과물을 생성한다. 학습자는 이를 통해 텍스트로만 존재하던 문학 작품을 시각적·청각적으로 재해석하고, 자신의 해석을 담은 디지털 콘텐츠를 직접 창작하는 경험을 획득한다. 이는 문학 작품에 대한 깊은 몰입을 유도하고 창의적 감상 능력을 배양하는 효과적 방법론이 될 수 있다.을 유도하고 창의적인 감상 능력을 길러주는 효과적인 방법이 될 수 있다.



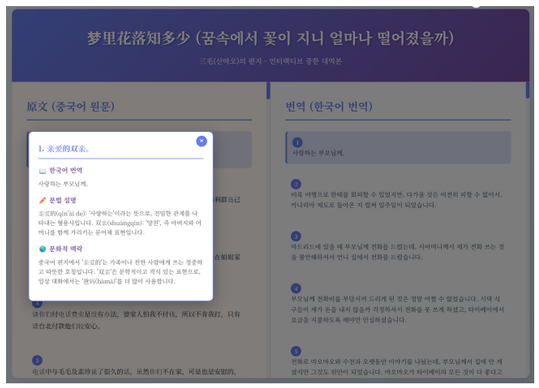
6) New York Times 'Hard Fork' Podcast, 2025.9.5.

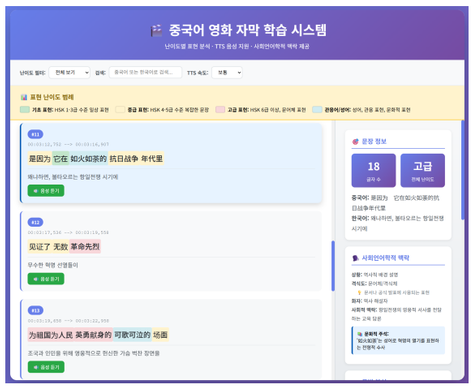
외국어 학습 자원의 웹앱 시각화

외국어 교육 영역에서 바이브 코딩은 학습자가 언어별 특성을 반영한 맞춤형 인터랙티브 학습 웹앱을 직접 제작할 수 있는 혁신적 도구로 기능한다. 전통적 교재 중심의 수동적 학습에서 벗어나, 학습자는 자신의 필요와 관심사에 부합하는 학습 자원을 능동적으로 창출하는 주체가 된다.

예를 들어, “외국어 학습자가 출발어 도착어, 유형, 난이도, 길이 등을 설정하여 외국어 통번역 학습을 하고 실시간 피드백하는 웹앱을 제작하라”고 AI에게 지시할 수 있다. AI는 음성 인식과 시각화를 결합한 애플리케이션을 즉시 생성하며, 학습자는 해당 언어의 통번역과 실시간 피드백 경험을 얻는다.

문학적 텍스트를 배우는 학습자는 문학 작품을 한국어로 번역하면서 외국어 TTS, 문법 설명과 문화적 맥락이 팝업으로 제시되는 인터랙티브 웹페이지를 제작할 수 있다[7]. 또 다른 학습자는 격식 체계와 사회적 위계가 중요한 언어를 대상으로, 애니메이션이나 대화 장면 속 표현을 격식 수준별 색상 자막으로 표시하고 사회언어학적 맥락을 제공하는 인터랙티브 웹페이지를 만들 수 있다.





7) http://claude.ai/public/artifacts/2da81f71-625f-4005-9f98-0a8631132e18

5) AI 시대 융합 교육을 위한 교수자 역량의 재정의

교육 자료의 기획, 준비, 서비스 역량의 재구성

AI 융합 인문학 교육 환경에서 교수자의 역할은 전통적 지식 전달자에서 '학습 경험 설계자(Learning Experience Designer)'로 전환되어야 한다. 개인 맞춤형 AI 교육 앱을 활용하는 '알파 스쿨(Alpha School)'에서는 교사를 '가이드(Guide)'라 칭하며, 이들은 정보 전달보다 학습자에게 동기를 부여하는 역할에 주력한다. 이처럼 교수자는 더 이상 정제된 지식을 일방적으로 전달하는 데 그치지 않고, 학습자가 AI와 상호작용하며 자율적으로 지식을 탐구하고 구성해 나갈 수 있는 학습 환경과 프로젝트를 설계해야 한다. 교육자들은 수업 설계, 학생 상담과 같이 창의성이나 직접적 상호작용이 필요한 업무에는 AI를 보완 수단으로 활용하고, 행정 업무와 같은 반복적 작업은 자동화하는 경향을 보인다[8]. 즉, 교수자는 교육 콘텐츠 기획자, 프로젝트 매니저, 그리고 서비스 운영자의 역할을 복합적으로 수행해야 한다.

프롬프트 엔지니어(Prompt Engineer)로서의 교수자

특히 생성형 AI를 교육에 효과적으로 활용하기 위해서는 '프롬프트 엔지니어링(Prompt Engineering)' 능력이 필수적이다. 프롬프트는 AI로부터 원하는 결과물을 도출하기 위해 입력하는 질문이나 명령어를 의미하며, 프롬프트 설계 방식에 따라 AI가 생성하는 결과물의 질이 크게 좌우된다. 융합 교육에서 교수자는 학습자가 특정 학습 목표를 달성하는 데 최적화된 '교육용 프롬프트'를 개발하고 제시하는 역할을 수행해야 한다. 예를 들어, 단순히 "소크라테스에 대해 설명하라"는 프롬프트 대신, "당신은 소크라테스가 되어 문답법(Socratic Method) 스타일로 학습자인 나와 대화하며 '정의'의 개념에 대해 탐구하라. 먼저 나에게 '정의란 무엇이라고 생각하는가?'라는 질문을 제기하라"와 같이 구체적 역할, 맥락, 상호작용 방식을 명시하는 정교한 프롬프트를 제공할 수 있다. 이러한 프롬프트 설계 능력은 AI를 효과적인 교육 파트너로 전환하는 핵심 역량이 될 것이다.

3. 결론: 공유와 확산을 통한 인문 융합 교육의 지속 가능한 미래

수퍼 프롬프트(Super Prompt)의 공유

AI 기반 융합 인문학 교육이 성공적으로 안착하고 확산되기 위해서는 개별 교수자의 노력을 넘어, 학문 공동체 차원의 협력과 자원 공유가 필수적이다. 특히 각 인문학 전공 영역(역사학, 철학, 문학, 언어학 등)의 고유한 특성에 맞춰 개발된 효과적 교육 콘텐츠, 웹 애플리케이션 개발 사례, 그리고 무엇보다 양질의 결과물을 생성하도록 유도하는 "수퍼 프롬프트(Super Prompt)"를 체계적으로 수집하고 공유하는 플랫폼 구축이 요구된다. '수퍼 프롬프트'란 특정 교육 목표 달성을 위해 다각도로 정교하게 설계된 프롬프트의 집합체를 의미한다.

예를 들어, '역사학 수업을 위한 사료 비판 수퍼 프롬프트', '문학 비평문 작성을 위한 AI 브레인스토밍 파트너 수퍼 프롬프트' 등 각 전공의 방법론과 교육 목표가 내재된 프롬프트 템플릿을 개발하고 공유하는 것이다. 이러한 공유 시스템은 교수자가 융합 교육을 준비하는 데 소요되는 시간과 노력을 획기적으로 절감할 뿐만 아니라, 집단 지성을 통해 교육의 질을 지속적으로 개선해 나가는 선순환 구조를 창출할 수 있다.

지속 가능한 인문학 교육 생태계 구축을 향하여

이제 교육계가 제기해야 할 질문은 'AI를 어떻게 활용할 것인가'가 아니라 '교육 자체를 어떻게 근본적으로

8) Anthropic 교육 보고서, 2025. 8. 27.

재구조화할 것인가'이다. AI와 인문학의 융합은 더 이상 선택사항이 아닌 필수불가결한 과제이며, 우리는 기술의 도전에 수동적으로 대응하는 것을 넘어, 기술을 적극적으로 활용하여 인문학 교육의 본질을 강화하고 새로운 가능성을 탐색해야 한다.

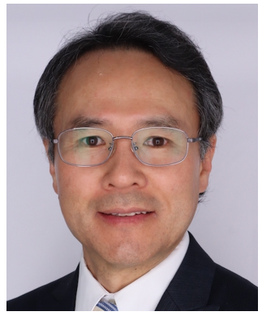
코딩 없는 융합 교육 환경은 기술 장벽을 허물고 더 많은 인문학 전공자가 자신의 지적 탐구 결과를 창의적으로 표현하고 세계와 공유할 수 있는 기회를 제공한다. 한국외국어대학교의 사례에서 확인할 수 있듯, '사유하는 인문학'은 이제 '보고 공유하는 인문학'으로 진화하고 있다.

이를 실현하기 위해서는 교육과정의 혁신, 교수자 역량의 강화, 그리고 학문 공동체 차원의 긴밀한 협력이 필요하다. AI 기술을 통해 학습자의 능동적 참여를 극대화하고, 교수자는 효과적인 교육 경험 설계자이자 프롬프트 엔지니어로 재탄생해야 한다. 또한, 성공적 교육 모델과 수퍼 프롬프트를 공유하고 확산함으로써 지속 가능한 인문 융합 교육 생태계를 협력적으로 구축해 나가야 할 것이다.

이러한 노력을 통해 인문학은 AI 시대의 위기를 초월하여, 인간과 기술이 조화롭게 공존하는 미래 사회를 선도하는 핵심 학문으로 재정립될 수 있을 것이다. 인문학의 미래는 기술과의 대립이 아닌 융합에 있으며, 이는 곧 인간 존재의 의미와 가치를 더욱 깊이 있게 탐구하는 새로운 인문학 르네상스의 시작이 될 것이다.

협력적 사회 생산에서의 인공지능: 현재의 전개와 미래 의제

Artificial Intelligence in Collaborative Social Production: Present Developments and Forward-Looking Agendas



강석
텍사스대학교 교수

Seok Kang
Professor, The University of Texas at San Antonio

Abstract

This paper aims to qualitatively synthesize the role of AI in social production through thematizing scholarly works in the humanities and social sciences. Artificial Intelligence's (AI) role in the theoretical and practical aspects of philosophy, history, culture, language, art, music, ethics, and communication was elaborated, and evidence-based information was provided. Three major themes: human-AI collaboration, ethical and epistemological influence of AI, and AI as a transdisciplinary catalyst represented research evidence of AI in the humanities and social sciences literature. Based on the findings, this paper proposes the synthetic human-AI collaborative social production model, in which human-AI collaborations facilitate knowledge generation, application, and implementation. AI as an extension of the human mind suggests the key role of the humanities and social sciences in the coexistence of AI in the human world. Forward-looking agendas in social production with AI were detailed.

Keywords: Artificial intelligence, Social production, Human-AI collaboration, Interdisciplinary coexistence of humans and technology

A paradigmatic shift in the Fourth Industrial Revolution has been accelerated by artificial intelligence (AI). Landing in the lay public's daily routines through generative AI (GenAI), a type of AI creating content through machine learning, deep learning, and reinforcement learning, such as ChatGPT, Copilot, Perplexity, and Midjourney, accessibility and applicability have been secured and are in continuous progress. Both hardware and software industries introduce new AI products and services to the market, which permeate the fabric of people's digital lives. A resultant phenomenon is data-driven social production (Burns, 2025). Social production refers to the process of creating, organizing, and distributing tangible goods, intangible services, knowledge, and cultural products by social members in communicative and relational dimensions at the societal level (Kiggins, 2021). In the AI context, social production indicates the collaborative creation of cultural, knowledge, and symbolic works in social relations and structures using AI. From the humanistic and social scientific AI perspectives, social members are creators, consumers, mediators, definers, collaborators, and producers of narratives, identities, ideas, and meanings.

A new norm has been created as AI is used in social production in the global academia and industries. Social outcomes through production convene with theoretical insights and applications. Social production through AI is an interdisciplinary combination of engineering, science, humanities, and social sciences (Priya et al., 2025). Feasible deliverables in the lens of social production with AI imply discussions of human labor, culture, ethics, and communication because outputs operate and are used contextually. High-caliber studies on social production with AI in science, technology, engineering, and mathematics (STEM) are widely available (e.g., Channi et al., 2025; Rzevski, 2025). Scholarly works in the humanities and social sciences (HSS) emphasize the catalytic role of AI for encouraging social enhancement and mitigating risks (e.g., Bozkurt & Gursoy, 2025; Karjus, 2025).

Opportunities and challenges on the verge of social production with AI coexist. With AI, individuals and communities can access GenAI tools to co-create content (Rizun et al., 2025). AI also enriches knowledge production experiences as it assists in writing and summarizing public discourse (Maci & Anesa, 2025). Once ideas are designed, collective intelligence can be built through AI. In the process, broad participation empowers community engagement and abates bias because social producers collectively monitor the outcomes (Kannan, 2025).

Meanwhile, challenges exist on the other side of the coin. An AI divide can occur between societies with a multitude of large tech companies and the countries with limited access (Wang et al., 2025). Data can be controlled by AI if users depend on what the tools produce. Amid the control, cultural and social biases can linger (Jenks, 2025). The role of human agents remains a question as AI dominates in social production. Accountability and ethical transparency can be neglected by AI-driven ecosystems because AI-generated content becomes countless and exceeds available fact-checking capacities (Monnier & Ségur, 2025). On top of AI dominance, a question of human roles in coexistence with AI arises.

In discussing insights into the present and future of AI in social production, some limitations of past studies can be addressed. While the adoption, creation, and production of social deliverables with AI are underway, a critical and synthetic view on the role of HSS in AI development is scarce. STEM's impact on AI production is well-received and continues to shape society. However, academic papers on the contemporary form of social production with AI from HSS perspectives are understudied. As a normative science, HSS enables leveraging critical thinking, philosophical, analytical, and storytelling competencies for ethical, social, and cultural AI outputs (Olojede & Polo, 2025). A qualitative review of HSS studies on AI can provide a synthetic perspective on social production, which will facilitate collaborations, discover the crucial role HSS plays in the AI ecosystem, and contribute to laying out action plans for the coexistence of humans and technology by building a global consensus.

This paper aims to qualitatively synthesize the role of AI in social production through thematizing scholarly works in HSS. In the current paper, AI's role in the theoretical and practical aspects of philosophy, history, culture, language, art, music, ethics, and communication is elaborated, and evidence-based information is provided. An interpretation from a thematic analysis can offer a basis for suggesting forward-looking agendas in social production with AI from HSS standpoints.

AI and Social Production in the Humanities and Social Sciences

As a paradigmatic shift, AI not only alters the way humans use technology but also shapes society because AI is a general-purpose technology disrupting social relations (Kiggins, 2021). Social production with AI refers to the shaping of social structures and practices through AI technologies (Padiyath et al., 2024). Due to the expansive role of AI in the social fabric, the concept of collaborative social production emerges. Using AI, humans co-produce feasible deliverables with AI in human-only, AI-only, modified-AI, and AI-guided modalities. In a study on the comparative analysis of modal effects, AI-guided messages were rated as more authentic and helpful than the other modalities (Meng et al., 2025).

Known as AI-Mediated Communication (AI-MC), human-AI collaborations modify, augment, and generate messages to accomplish communicative goals at micro, meso, and macro levels. Human-AI collaborations for social production involve generating ideas, negotiating conflict resolutions, selecting the most compelling plan, and implementing the designed proposal (Joshi, 2025). Message generation patterns shape narrative features, resulting in audiences' evaluation of received communication. Optimal match theory (Cutrona & Russell, 1990) applied to the social production with AI context suggests that the modified-AI

modality generates informational and emotional narratives, and the match between humans and AI positively influences authenticity evaluations (Meng et al., 2025). Therefore, human-AI collaborative social production assumes that the ideal and fact-checked match between humans and AI in generating feasible deliverables leads to positive message authenticity, attitude, and behaviors. In turn, social production with AI is more likely to be supplementary to, rather than a replacement for, human activities.

HSS are pivotal venues of human-AI collaboration in terms of social production. Philosophical perspectives are embedded in ethical principles that reflect human moral reasoning (Kumar, 2025). AI should make decisions that align with societal values. As another example, the history discipline in HSS is properly positioned in the discussion of social production with AI by refining fact-checked history. Human-AI collaboration can account for immersive experiences generated by AI in historical narratives (Manitsaris et al., 2025). AI-driven speech data analytics enables language learning and pedagogical development (Bai & Nordin, 2025). In an interaction between humans and AI, the experience increases the possibility of political interests and voter turnout (Šola, et al., 2025). Informed choices and effective communication are made when AI-guided messages are authenticated (Mseer et al., 2025). On the flip side, AI produces deceptive results, raising concerns about ethics and credibility (e.g., Patama, 2025).

Collecting numerous data and combining them collectively, AI produces subsequent results on its own. Starting with human inquiries, AI can guide through the information generation and decision-making processes. This is where human-AI collaboration activities come into play for perceptual, attitudinal, and behavioral outcomes.

AI's Role in Social Production: Trans-Disciplinary Deliverables

HSS provides a roadmap for social production to occur and be evaluated within the AI adoption process. Technology, including AI, is actively shaped by social, cultural, and political factors. AI is a reciprocal technology that emphasizes human agency and contextualized development for the social world (Gretzky & Dishon, 2025). Philosophy, History, Linguistics, Politics, Art, Music, and Communication disciplines in HSS provide rich research evidence addressing AI's role in related social production.

In philosophy, the fundamental question about AI is whether AI systems know the knowledge or merely yield answers from natural language data processing (Pakarinen & Huising, 2025). AI is embedded in the relational expertise that creates meaningful inroads into non-routine reasoning about intricate cases. That expertise is more accurately conceptualized as relationally constituted knowledge through interactions. In this domain of relational expertise, philosophers debate the bias, trust, and epistemic authority of AI-generated knowledge (Mahbubi, 2025). Research on AI in philosophy focuses on the epistemological relationship between AI and humans by aiming at balancing technological capabilities with ethical considerations.

As such, the key contribution of philosophy to AI is ethics (Müller, 2025). Philosophers explore fairness, accountability, transparency, privacy, responsibility, and human values. Consensus, conflicts, and synthesis between humans and AI from the perspectives of embodiment, consciousness, sentience, and stimulation reflect philosophical works in the AI domain (Pauketat, 2025). Additionally, governance, equity, aesthetic value, and collaborative creation between humans and machines are the philosophical agendas of social production with AI (Mahajan, 2025). Hence, reciprocal interactions between humans and AI are the basis of philosophical discussions.

History studies pay special attention to AI in historical knowledge production and curation. AI as a co-agent accelerates the interpretation of historical texts. Ethical challenges posed by AI are a concern in historical research as well. Historians raise the issues of transparency, bias, and potential misrepresentation requiring ongoing attention (Kaul, 2025). AI-powered public history and education in the form of interactive exhibits, virtual reality storytelling, narrative construction, and chatbots in museums illustrate human-AI dynamics (Han, 2025).

AI studies and applications in HSS continue to evolve. The linguistics discipline is directly affected by AI as well. Using natural language processing (NLP), AI models analyze language structures (Davenport, 2025). Word formation and inflections are anatomized and patterned through AI morphology (Pascual-Triana et al., 2025). Speech recognition analyzes sound patterns and helps with pronunciation corrections in phonetics. Linguists can conduct topic modeling and clustering to identify themes from a corpus of textual data (Kirilenko & Stepchenkova, 2025). Language tracking and sentiment analyses are frequent research examinations. Practically, AI tools are used for speech-to-text conversion and translation (Rouf & Jadon, 2025). To foreign language learners, AI tools have become a key platform.

Political science is undergoing transformations in theory and practice through social production with AI. As an exponentially increasing number of researchers and politicians adopt AI, scholars explore how AI shapes democracy, power, and deliberation in algorithm-based decision-making. Machine-learning simulations and AI modeling are used to test political science theories (Nielsen & Zhou, 2025). With elaborations provided by the procedures, AI theorists predict policy impacts and political engagement in the public. AI is used as a collaborator rather than a tool in political knowledge framework production (de Slegte et al., 2025). In this process, questions about content authenticity, bias, and interpretability arise. On the practical side, professionals and government agents use AI

to analyze policies and forecast future outlooks. AI enables political campaigners to learn the patterns of speeches, voter turnout, manifestos, and pledges, resulting in helping campaign strategies and public opinion tracking (Kumar et al., 2025). Political strategies, public mobilization, and disinformation detection can be designed and implemented using AI. Concerns about disillusionment and responsible AI in political science emerge as well. These findings suggest that human-AI collaborations are synergistic rather than a zero-sum game in social production with AI in politics.

Digital platforms and tools equipped with AI technology contribute to artistic production. In art, AI is a non-human actor that actively shapes artistic outcomes by interacting with human actors, according to actor-network theory (Morton, 2025). Creativity generated by AI is not a standalone entity but an embedded agency in the human-machine environment. One caveat is the ownership of socially produced artwork with AI. In that sense, intellectual property discussions on creative collaborations with AI are an ongoing debate (Hwang et al., 2025). As co-creation between artists and AI becomes a frequent occasion, social production in the art discipline blends new media studies, ethics, and aesthetics. Using DALL-E, RunwayML, or Magenta, for example, AI-guided artworks enable community-driven projects. As an upside, AI can lower entry barriers for marginalized creators, shifting social production of artworks with AI from select artists to community members (Oppenlaender et al., 2025). On the other hand, disinformation with malicious intentions embodied in artwork can generate authenticity, ethics, and verification issues (Khatiwada et al., 2025).

In music, the collaborative creation, performance, and distribution of musical works with AI is becoming a new normal. From a theoretical standpoint, actor-network theory treats generative AI, algorithms, and recommendation systems as actors, being part of a network of human and non-human interactions (Väkevä & Partti, 2025). AI fosters a participatory culture in music production because AI tools, such as Soundtrap, Amper Music, or AIVA, open a social production forum where users from diverse backgrounds co-create music works. AI music tools can invoke issues of originality, human-machine labor divisions, and human musicians' livelihood (Sahoo, 2025). In the practical domain, AI-assisted composition is becoming prevailing (Tchemeube et al., 2025). From AI-guided music composition, human musicians benefit from social music production. Furthermore, live performance with AI systems, including Shimon, influences music experiences. AI's unlimited capabilities of music creation pose questions of musicians' identity, musical bias, and cultural reshaping (Ansani et al., 2025).

The communication discipline is at the epicenter of social production with AI. Mediated communication occurs between humans and AI. AI-only, AI-guided, or human-only interactions induce differing perceptual, attitudinal, and behavioral responses (Meng et al., 2025). AI contributes to content generation and decision-making. Hence, a new public sphere is created in AI. Algorithms and generative models shape public discourse, which can influence social engagement, democracy, and collective social production (Misnikov & Samoilava, 2025). As a part of a distributed agency, social production with AI becomes a collaborative process through the engagement of humans, machines, and systems. The influence of AI-generated signs, symbols, narratives, and visuals can be either beneficial or detrimental (Park et al., 2024). Media ecology and cultural communication prompt social production narratives with AI (Petricini, 2024).

Multiple AI tools permeate communication-related fields. For example, journalists use AI for news article production and fact-checking (Cazzamatta & Sarısakaloğlu, 2025). AI is utilized for community engagement projects and campaigns. Intercultural communication is facilitated by AI using captions, translations, mutual understanding, and cultural engagement (Sarwari et t al., 2024). Human-AI collaborations for communication synergize productive outcomes. However, misinformation or hallucination from AI can lead to public deception and ethical concerns (Saeidnia et al., 2025).

Themes from Social Production with AI in HSS

A synthetic look at theoretical developments and practical applications of AI in HSS implies that insightful approaches to AI, in addition to technical production, are an inevitable pillar of academic paradigms. Critical views and interpretations in HSS connect the missing link between humans and AI that STEM can overlook. From philosophy to art to communication, humanistic AI touts mediated co-creation, ideation, and implementation. Given the review of AI-related academic resources in HSS, three major themes can be identified from disciplinary works on social production with AI. A qualitative synthesis from the previous section shows a clear pattern. A qualitative synthesis analysis enables a contextual understanding of a phenomenon (Baxter et al., 2012).

Human-AI Collaboration. AI is a collaborative partner in social production. Human-AI interactions enhance idea development, design, and completion (Edwards et al., 2025). AI-guided communication yields synergy, where human creativity meets AI's prompt responses to outcomes. As seen in this theme, AI is an active shaper of social production. Possibly, future society can be an AI-embedded environment as co-creation becomes a norm. HSS will play a central role in the collaboration process.

Ethical and Epistemological Influence of AI. As a challenge, social shaping of AI technology raises a question of AI-generated knowledge, cognition, emotion, and discourse. AI's epistemic legitimacy and ethical principles are an ongoing debate in philosophy. Transparency and accountability are significant agendas that require AI literacy (Tsarouhas & Grigoriadis,

2025). HSS needs to be in the key position to answer critical questions about ethics, bias, trust, and moral responsibilities.

AI as a Transdisciplinary Catalyst. An innovative forum is laid out by AI technology. In the forum, AI enhances interdisciplinary research between HSS and STEM or among HSS disciplines by supplementing mutually lacking points (Srivastava et al., 2025). For example, public discourse analysis with communication expertise can be elaborated with AI linguistics. Social production with AI in art and music contributes to health communication through therapeutic implementation.

From the themes, the following diagram accounts for the reciprocal linkage among collaboration, ethics, and interdisciplinarity (Figure 1). AI ethics works as the foundation for sound human-AI collaborations in HSS. Interdisciplinary supplements can incorporate the power of HSS into insights, contributions, and social production.

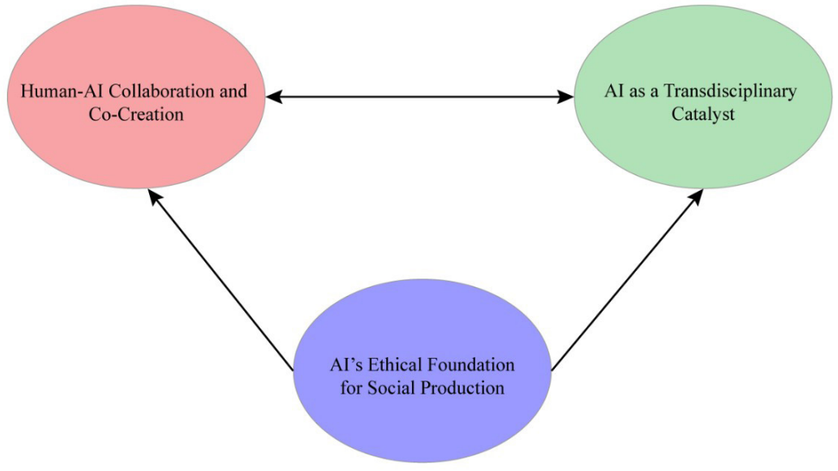


Figure 1. Three Themes of Social Production with AI in the Humanities and Social Sciences

Conclusion

This paper conceptualized social production with AI and applied it to HSS. As evident in the review and generated themes, the prevailing perspective centers on the interaction between humans and AI. HSS helps build ethical and philosophical underpinnings in the human-centered technology environment. HSS provides insights into cultural and social contextualization in AI's social production. Through collaborations, AI practices are produced as social outcomes. Future users are **AI**dience (AI + audience) who require robust literacy skills to be competent participants in the human-AI ecosystem. In this paper, the author proposes the **S**ynthetic human-**A**I collaborative social production **M**odel (SAM). SAM posits that human-AI collaborations lead to positive perceptual, attitudinal, and behavioral outcomes. AI-guided social production extends human-only cognitive processes by providing augmented neural-emotional integration. AI transcends its role as a mere tool and joins human minds as an external cognitive component, facilitating humans' decision-making capacity (Molina et al., 2024). This process is a human-AI co-evolution, which involves critical thinking, ethics, culture, and policy (Figure 2; AI-generated based on data input). HSS is at the epicenter of the coexistence framework.

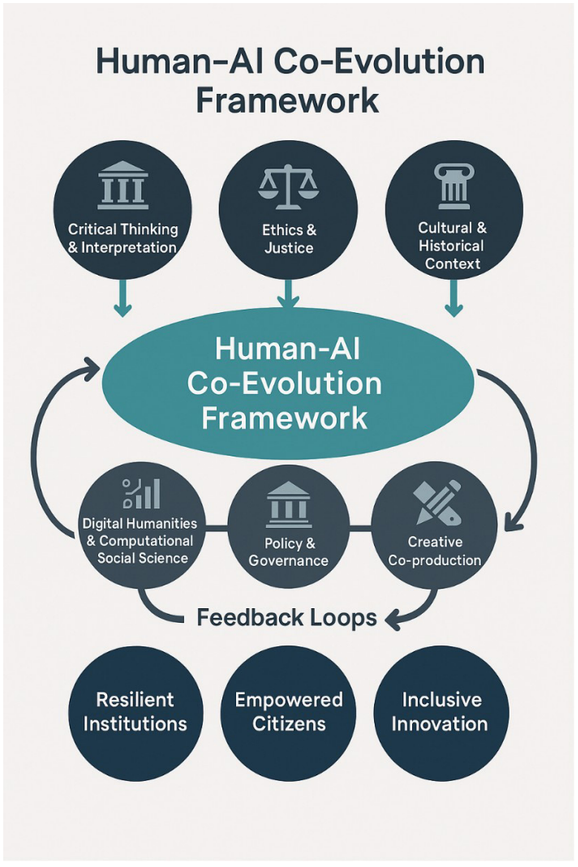


Figure 2. SAM Structure (AI-Generated Diagram based on data input)

This paper's evaluations of the current status of AI in terms of its coexistence with humans in HSS suggest several forward-looking agendas. They are promising viewpoints, as strengths will be bolstered and weaknesses will be amended. First, HSS needs to focus on human-centered AI. Robust AI values the solid deep learning pipeline. Humans' central role in collaborating, evaluating, and enhancing AI robustness is well-received and will be advocated (Tocchetti et al., 2025). Second, interdisciplinary AI research shapes an academic culture. HSS plays a key role in this practice because ethical, social positivist, and interpretive

perspectives build theoretical foundations. Interdisciplinary approaches can address contemporary social and technological challenges by fostering a deeper understanding of AI principles. Without sociocultural implications, AI research limits holistic assessments (Segessenmann et al., 2025). Third, global efforts to establish AI's ethical governance are suggested. With AI's dominance in the future world, a standardized manifesto for ethical social production and use of AI is in demand. From entertainment to news to disinformation, AI can readily manipulate agendas. At multiple levels, from team to organizational to national to international, who is governing, what is being governed, when it is being governed, and how it is being governed can provide guidance for ethical use. HSS can lead this effort (Batool et al., 2025).

Fourth, awareness needs to be put on AI divides. Age group discrepancies and national asymmetry in AI infrastructure, access, skill sets, and education opportunities exist, which can cause the social participation divide (Wang et al., 2025). HSS can be leading disciplines in narrowing the divide through research, practical social projects, community engagement, and programs. In this view, AI literacy is the fifth forward-looking agenda. By the definition of AI literacy, users consume, evaluate, and apply AI tools to participate in the digital world. Firm AI literacy paves the road to AI self-efficacy, the belief in success at a task (Bewersdorff et al., 2025), indicating that HSS needs to take the steering role in conducting AI literacy programs.

In conclusion, human-AI collaborative social production is human-centered across multiple disciplines with ethical governance. With efforts to overcome divides and provide literacy programs, HSS coexists with AI in human lives. This paradigmatic shift in the shaping of society with technology, extended human minds, and AI-guided social production suggests what HSS can do in the transformation.

References

Ansani, A., Koehler, F., Giombini, L., Hämäläinen, M., Meng, C., Marini, M., & Saarikallio, S. (2025). AI performer bias: Listeners like music less when they think it was performed by an AI. Empirical Studies of the Arts, 43(2), 1137-1161. https://doi.org/10.1177/02762374241308807 (Original work published 2025)

Bai, X., & Nordin, N. R. M. (2025). Human-AI collaborative feedback in improving EFL writing performance: An analysis based on natural language processing technology. Eurasian Journal of Applied Linguistics, 11(1), 1-19.

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. AI and Ethics, 5, 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

Baxter, S., Enderby, P., Evans, P., & Judge, S. (2012). Barriers and facilitators to the use of high-technology augmentative and alternative communication devices: A systematic review and qualitative synthesis. International Journal of Language & Communication Disorders, 47(2), 115–129. https://doi.org/10.1111/j.1460-6984.2011.00090.x

Bewersdorff, A., Hornberger, M., Nerdel, C., & Schiff, D. S. (2025). AI advocates and cautious critics: How AI attitudes, AI interest, use of AI, and AI literacy build university students' AI self-efficacy. Computers and Education: Artificial Intelligence, 8, 100340. https://doi.org/10.1016/j.caeai.2024.100340

Bozkurt, V., & Gursoy, D. (2025). The artificial intelligence paradox: Opportunity or threat for humanity? International Journal of Human–Computer Interaction, 41(1), 174–187. https://doi.org/10.1080/10447318.2023.2297114

Burns, K. S. (2025). Artificial intelligence: Helping or harming the creative spirit. In J. Costello & S. Yesiloglu (eds.), Influencer Marketing: Building Brand Communities and Engagement (pp. 227-248). Routledge. https://doi.org/10.4324/9781003434498-15

Cazzamatta, R., & Sarısakaloğlu, A. (2025). Mapping global emerging scholarly research and practices of AI-supported fact-checking tools in journalism. Journalism Practice, 1–23. https://doi.org/10.1080/17512786.2025.2463470

Channi, H. K., Kaur, A., & Kaur, S. (2025). AI-driven generative design redefines the engineering process. Generative Artificial Intelligence in Finance: Large Language Models, Interfaces, and Industry Use Cases to Transform Accounting and Finance Processes, 327-359. https://doi.org/10.1002/9781394271078.ch17

Cutrona, C. E., & Russell, D. W. (1990). Type of social support and specific stress: Toward a theory of optimal matching. In B. R. Sarason, I. G. Sarason, & G. R. Pierce (Eds.), Social support: An interactional view (pp. 319–366). John Wiley & Sons.

Davenport, M. J. (2025). Enhancing legal document analysis with large language models: A structured approach to accuracy, context preservation, and risk mitigation. Open Journal of Modern Linguistics, 15(2), 232-280. https://doi.org/10.4236/ojml.2025.152016

de Slegte, J., Van Droogenbroeck, F., Spruyt, B., Verboven, S., & Ginis, V. (2025). The use of machine learning methods in political science: An in-depth literature review. Political Studies Review, 23(3), 764-784. https://doi.org/10.1177/14789299241265084 (Original work published 2025)

Edwards, J., Nguyen, A., Lämsä, J., Sobocinski, M., Whitehead, R., Dang, B., ... & Järvelä, S. (2025). Human-AI collaboration: Designing artificial agents to facilitate socially shared regulation among learners. British Journal of Educational Technology, 56(2), 712-733. https://doi.org/10.1111/bjet.13534

Gretzky, M., & Dishon, G. (2025). Algorithmic-authors in academia: blurring the boundaries of human and machine knowledge production. Learning, Media and Technology, 1–14. https://doi.org/10.1080/17439884.2025.2452196

Han, W. (2025). Study on interactive experience design of AI in digital display of museums. International Journal of High Speed Electronics and Systems, 2540658. https://doi.org/10.1142/S0129156425406588

Hwang, Y., Shin, D., & Lee, J. H. (2025). Who owns AI-generated artwork? Revisiting the work of generative

AI based on human-AI co-creation. Telematics and Informatics, 98, 102266. https://doi.org/10.1016/j.tele.2025.102266
Jenks, C. J. (2025). Communicating the cultural other: Trust and bias in generative AI and large language models. Applied Linguistics Review, 16(2), 787-795. https://doi.org/10.1515/applirev-2024-0196
Joshi, S. (2025). Review of autonomous and collaborative agentic AI and multi-agent systems for enterprise applications. International Journal of Innovative Research in Engineering and Management 12(3), 65-76.
Kannan, S. (2025). Transforming community engagement with generative AI: Harnessing machine learning and neural networks for hunger alleviation and global food security. Cuestiones de Fisioterapia, 54(4), 953-963.
Karjus, A. (2025). Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. Humanities and Social Sciences Communications, 12, 277.
Kaul, A. (2025). Systematic review of literature ethics and role of ai in historical research and reconstruction. Gap Bodhi Taru, 1-6.
Khatiwada, P., Washington, J., Walsh, T., Hamed, A. S., & Bhatta, L. (2025). The ethical implications of AI in creative industries: A focus on AI-generated art. arXiv preprint arXiv:2507.05549. https://doi.org/10.48550/arXiv.2507.05549
Kiggins, R. D. (2021). Social production and artificial intelligence. In T. Keskin & R. D. Kiggins (Eds.), Towards an international political economy of artificial intelligence (pp. 3–16). Cham, Switzerland: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-74420-5_1
Kirilenko, A. P., & Stepchenkova, S. (2025). Facilitating topic modeling in tourism research: Comprehensive comparison of new AI technologies. Tourism Management, 106, 105007. https://doi-org.libweb.lib.utsa.edu/10.1016/j.tourman.2024.105007
Kumar, M., Singh, A. K., & Das, S. (2025). Sentiment analysis of political leaders speeches using AI and NLP with machine learning, deep learning, and transformer models. Open Access International Journal of Science & Engineering, 8(2), 32-38.
Kumar, S. (2025). Algorithmic Ethics and the Human Mind: A Cross-Disciplinary Study on AI Decision-Making and Moral Philosophy. Bookgram Publishers Multidisciplinary Academia Journal p-ISSN 3051-2379 e-ISSN 3051-2387, 1(01), 1-9.
Maci, S. M., & Anesa, P. (2025). The impact of AI on discourse analysis: Challenges and opportunities. International Journal of Language Studies, 19(2).
Mahajan, P. (2025). The soul of the AI: Governance, ethics, and the future of human–AI integration. Zenodo. https://doi.org/10.5281/zenodo.15789678
Mahbubi, M. (2025). Digital Epistemology: Evaluating the credibility of knowledge generated by AI. YUDHISTIRA: Journal of Philosophy, 1(1), 8-18.
Manitsaris, S., Glushkova, A., Spyridonos, A., & Senteri, G. (2025). Working with AI in cultural and creative industries. Presse des Mines. https://hal.science/hal-05016368
Meng, J., Zhang, R., Qin, J., Lee, Y.-J., & Lee, Y.-C. (2025). AI-mediated social support: The prospect of human–AI collaboration. Journal of Computer-Mediated Communication, 30(4), zmaf013. https://doi.org/10.1093/jcmc/zmaf013
Misnikov, Y., & Samoilava, V. (2025, June). Building an AI-Supported public discourse model. In 2025 Eleventh International Conference on eDemocracy & eGovernment (ICEDEG) (pp. 286-291). IEEE. https://doi.org/10.1109/ICEDEG65568.2025.11081665
Molina, D. A., Kharlov, V., & Chen, J.-S. (2024). Towards effective human-AI collaboration in decision-making: A comprehensive review and conceptual framework. 2024 Portland International Conference on Management of Engineering and Technology (PICMET), 1–6. https://doi.org/10.23919/PICMET64035.2024.10653303
Monnier, A., & Ségur, C. (2025). Challenges for Fact-checking: Beyond False/True Verification. InMedia. The French Journal of Media Studies, 10, 1-19.https://doi.org/10.4000/145d4
Morton, J. L. (2025). On inscription and bias: Data, actor network theory, and the social problems of text-to-image AI models. AI and Ethics, 5, 775–790. https://doi.org/10.1007/s43681-024-00431-8
Mseer, I., & Ali Samhan, A. A. (2025). Collaboration between humans and AI. In In: Alaali, M., Musleh Al-Sartawi, A.M.A., Aydiner, A.S. (eds). The Paradigm Shift from a Linear Economy to a Smart Circular Economy: The Role of Artificial Intelligence-Enabled Systems, Solutions and Legislations (pp. 1299-1308). Cham: Springer Nature Switzerland.
Müller, V. C. (2025). Philosophy of AI: A structured overview. A Companion to Applied Philosophy of AI, 14-30. https://doi.org/10.1002/9781394238651.ch2
Nielsen, R. A., & Zhou, A. Y. (2025). Integrating social science research across languages with assistance from artificial intelligence. Daedalus, 154(2), 51–67. https://doi.org/10.1162/daed_a_02140
Olojede, H. T., & Polo, E. P. (2025). In praise of normative science: Arts and humanities in the age of artificial intelligence. International Journal of Social Sciences and Humanities: Africa Research Corps Network (Arcn) Journals, 11(2), 1-9.
Oppenlaender, J., Johnston, H., Silvennoinen, J. M., & Barranha, H. (2025). Artworks reimagined: Exploring human-AI co-creation through body prompting. Proceedings of the ACM on Human-Computer Interaction, 9(4), 1-34. https://doi.org/10.1145/3734189
Padiyath, A., Hou, X., Pang, A., Viramontes Vargas, D., Gu, X., Nelson-Fromm, T., ... & Ericson, B. (2024, August). Insights from social shaping theory: The appropriation of large language models in an undergraduate programming course. In Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1 (pp. 114-130).
Pakarinen, P., & Huising, R. (2025). Relational expertise: What machines can't know. Journal of Management Studies, 62(5), 2053-2082. https://doi.org/10.1111/joms.12915
Park, S., Park, S., Kim, J., & Han, K. (2024, November). Exploring the impact of AI-generated images on political news perception and understanding. In Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (pp. 565-571). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3678884.3681907
Pascual-Triana, J. D., Fernández, A., Del Ser, J., & Herrera, F. (2025). Overlap number of balls model-agnostic counterfactuals (ONB-MACF): A data-morphology-based counterfactual generation method for trustworthy artificial intelligence. Information Sciences, 701, 121844. https://doi.org/10.1016/j.ins.2024.121844
Patama, S. (2025). Exploring Human-AI Collaboration in the Creative Process: Enhancements and Limitations. Master's Thesis. University of Jyväskylä.
Pauketat, J. V., Ladak, A., & Anthis, J. R. (2025). World-making for a future with sentient AI. British Journal of Social Psychology, 64(1), e12844. https://doi.org/10.1111/bjso.12844
Petricini, T. (2024). Special issue introduction: AI and media ecology. Explorations in Media Ecology, 23(2), 93-103. https://doi.org/10.1386/eme_00198_2
Priya, M., Anandan, V., Henrietta, M. H. M., & Varalakshmi, S. (2025). Recent Innovations in Sciences and Humanities. Taylor & Francis Group.
Rizun, N., Edelmann, N., Janowski, T., & Revina, A. (2025, May). AI-enabled co-creation for evidence-based policymaking: A conceptual model. In Conference on Digital Government Research (Vol. 1).
Rouf, M., & Jadon, J. S. (2025). Generative AI for text to speech and sign language translation. In 2025 3rd International Conference on Disruptive Technologies (ICDT) (pp. 1124–1129). IEEE. https://doi.org/10.1109/ICDT63985.2025.10986615
Rzevski, G. (2025). Artificial intelligence in engineering: past, present and future. WIT Transactions on Information and Communication Technologies, 10.
Saeidnia, H. R., Hosseini, E., Lund, B., Alipour Tehrani, M., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: A systematic review of challenges and approaches. Knowledge and Information Systems, 67, 3139–3158. https://doi.org/10.1007/s10115-024-02337-7

Sahoo, B., & Sakalkar, R. (2025). Music copyright in the age of artificial intelligence: A new era of creative ownership. Sangeet Galaxy, 14(2), 115–125.
Sarwari, A. Q., Javed, M. N., Mohd Adnan, H., & Abdul Wahab, M. N. (2024). Assessment of the impacts of artificial intelligence (AI) on intercultural communication among postgraduate students in a multicultural university environment. Scientific Reports, 14, 13849. https://doi.org/10.1038/s41598-024-63276-5
Segessenmann, J., Stadelmann, T., Davison, A., & Dürr, O. (2025). Assessing deep learning: A work program for the humanities in the age of artificial intelligence. AI and Ethics, 5, 1–32. https://doi.org/10.1007/s43681-023-00408-z
Šola, H. M., Qureshi, F. H., & Khawaja, S. (2025, March). Human-centred design meets AI-driven algorithms: Comparative analysis of political campaign branding in the Harris–Trump Presidential Campaigns. Informatics, 12(1), 30.
Srivastava, P., Choudhary, R. R., Tekwani, K., & d Srivastava, A. (2025). Implementation of AI in humanities. Recent Advances in Sciences, Engineering, Information Technology & Management. United States: CRC Press.
Tchemeube, R. B., Ens, J., Plut, C., Pasquier, P., Safi, M., Grabit, Y., & Rolland, J. B. (2025). Evaluating human-AI interaction via usability, user experience and acceptance measures for MMM-c: A creative AI system for music composition. arXiv preprint arXiv:2504.14071. https://doi.org/10.48550/arXiv.2504.14071
Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., & Yang, J. (2025). AI robustness: a human-centered perspective on technological challenges and opportunities. ACM Computing Surveys, 57(6), 1-38. https://doi.org/10.1145/3665926
Tsarouhas, P., & Grigoriadis, K. (2025, May). Building trust in AI for public administration: A strategic framework for transparency, XAI, participation, and digital literacy. In 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) (pp. 1-9). IEEE. https://doi.org/10.1109/ICHORA65333.2025.11017116
Väkevä, L., & Partti, H. (2025). Generative AI as a collaborator in music education: An action-network theoretical approach to fostering musical creativities. Action, Criticism, and Theory for Music Education, 24(3), 16-52. https://doi.org/10.22176/act24.3.16
Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & H de Vreese, C. (2025). The artificial intelligence divide: Who is the most vulnerable?. New Media & Society, 27(7), 3867-3889. https://doi.org/10.1177/14614448241232345