2025 국제 인권 콘퍼런스

International Human Rights Conference

신기술과 인권
인공지능의 기회와 도전

New Technology and Human Rights
Opportunities and Challenges of Artificial Intelligence

국가인권위원회

APF

발 간 등 록 번 호
11-1620000-100041-01

2025 국제 인권 콘퍼런스
International Human Rights Conference

신기술과 인권
인공지능의 기회와 도전

New Technology and Human Rights
Opportunities and Challenges of Artificial Intelligence

국가인권위원회

목차

개회식

세션 1. 신기술과 인권: 국제 동향과 국가별 대응

좌장: 키렌 피츠페트릭 | 아시아 태평양 국가인권기구포럼 사무국장

세션 2. 신기술과 인권 과제: 불평등과 차별의 문제

좌장: 김민호 | 성균관대학교 법학전문대학원 교수

목차

Contents

Contents

프로그램

시간	프로그램
09:00~09:30	등록
09:30~10:30	개회사 안창호 | 대한민국 국가인권위원회 위원장 축사 1 롤란트 호네캄프 | 주한 EU 대표부 차석대사 축사 2 사마르 알 하지 하산 | 아시아 태평양 국가인권기구포럼 의장, 요르단 국가인권기구 위원장 기조 발제 AI 발전 과정에 인권을 반영하기 필립 알스톤 | 뉴욕대학교 로스쿨 존 노톤 포메로이 교수, 전 유엔 빈곤과 인권 특별보고관
10:30~10:40	휴식
10:40~12:10	**세션 1. 신기술과 인권: 국제 동향과 국가별 대응** [좌 장] 키렌 피츠페트릭 | 아시아 태평양 국가인권기구포럼 사무국장 [발표자] 신기술과 인권: 글로벌 대응과 인권과제 백범석 | 경희대학교 법학전문대학원 교수, 유엔 인권이사회 자문위원회 위원 신뢰할 수 있는 AI와 기본권 평가의 교훈 로베르토 지커리 | 서울대학교 데이터사이언스대학원 특임교수 신기술과 인권 대응: 대한민국 AI 법제와 인권 보장을 위한 과제 안진현 | 국가인권위원회 인권정책과 사무관
12:10~13:30	점심
13:30~15:40	**세션 2. 신기술과 인권 과제 : 불평등과 차별의 문제** [좌 장] 김민호 | 성균관대학교 법학전문대학원 교수 [발표자] AI와 불평등의 재생산 이권일 | 경북대학교 법학전문대학원 교수 데이터, 사생활 보호, 그리고 차별 팀 엥겔하르트 | 유엔 인권최고대표사무소 인권담당관 기술 발전과 디지털 격차 나이갓 다드 | 디지털 권리 재단 상임이사, 유엔 AI 자문기구 위원 생존을 위한 시대, AI의 역할 마이클 키웻 | 요하네스버그대학교 사회변화연구센터 선임연구원, 예일대학교 방문 교수 디지털전환 시대 불안정노동과 AI 알고리즘 이승윤 | 중앙대학교 사회복지학과 교수
15:40~16:00	휴식

프로그램

시간	프로그램
16:00~17:30	**세션 3. 기술발전과 미래 그리고 대응** [좌 장] 안성율 | 국가인권위원회 정책교육국장 [발표자] AI 인권거버넌스: 인권기반 접근과 영향받는 사람 장여경 | 정보인권연구소 상임이사 AI 인권영향평가와 적용 유승익 | 명지대학교 법학과 객원교수 신기술과 인권에 대한 국가인권기구의 역할 넬레 루켄스 | 유럽 국가인권기구 연합 AI와 인권 실무그룹 의장
17:30~17:40	폐회식

Program

Time	Program
09:00~09:30	Registration
09:30~10:30	**Opening Remark** AHN Changho | Chairperson, National Human Rights Commission of Korea **Welcome Remark 1** Roland HONEKAMP | Deputy Chief of Mission, Delegation of the EU to the ROK **Welcome Remark 2** Samar Al Haj HASSAN | Chairperson, Asia Pacific Forum of National Human Rights Institutions Chairperson, Jordan National Centre for Human Rights **Keynote Speech** Making Human Rights Relevant in Shaping AI Philip ALSTON | John Norton Pomeroy Professor of Law, School of Law, New York University Former UN Special Rapporteur on Extreme Poverty and Human Rights
10:30~10:40	Coffee Break
10:40~12:10	**Session 1. New Technology and Human Rights: Global Trends and National Responses** [Moderator] Kieren FITZPATRICK | Director, Asia Pacific Forum of National Human Rights Institutions [Speakers] New and Emerging Technologies and Human Rights: Global Responses and Challenges BAEK Buhm-suk | Professor, School of Law, Kyunghee University Member, UN Human Rights Council Advisory Committee Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment Roberto V. ZICARI | Adjunct Professor, Graduate School of Data Science, Seoul National University Responding to Emerging Technologies and Human Rights: Tasks for Safeguarding Human Rights under Korea's AI Legal Framework AHN Jin-hyun | Deputy Director, Human Rights Policy Division, NHRCK
12:10~13:30	Lunch
13:30~15:40	**Session 2. Challenges of New Technologies: Inequality, Exclusion, and Human Rights** [Moderator] KIM Minho | Professor, School of Law, Sungkyunkwan University [Speakers] AI and the Reproduction of Inequality LEE Kwon il | Associate Professor, School of Law, Kyungpook National University Data, Privacy and Discrimination Tim ENGELHARDT | Human Rights Officer, OHCHR Technological Development and the Digital Divide Nighat DAD | Executive Director, Digital Rights Foundation Member, UN High-level Advisory Body on AI AI in the Age of Survival Michael KWET | Senior Researcher, University of Johannesburg Visiting Professor, Yale University Precarious work and AI Algorithms in the Digital Transformation Era LEE Seung yoon | Professor, Department of Social Welfare, Chung-Ang University
15:40~16:00	Coffee Break

Program

Time	Program
16:00~17:30	**Session 3. Responding to Human Rights Challenges in the Digital Age** [Moderator] AN Seong Ryul | Director General, Policy Bureau, NHRCK [Speakers] AI Human Rights Governance: Human Rights-Based Approach and Affected Persons CHANG Yeo-kyung | Executive Director, Institute for Digital Rights AI Human Rights Impact Assessment and Application YOO Seung-ik | Guest Professor, Department of Law, Myongji University The Role of NHRIs for the New Technology and Human Rights Nele ROEKENS | Chair of the Working Group on AI, ENNHRI
17:30~17:40	Closing Remark

신기술과 인권: 인공지능의 기회와 도전

New Technology and Human Rights: Opportunities and Challenges of Artificial Intelligence

개회사 & 축사

Opening Remark & Welcome Remarks

개회사 | Opening Remark

안창호

AHN Changho

대한민국 국가인권위원회 위원장

Chairperson, National Human Rights Commission of Korea

[주요경력]

- 제23회 사법시험 합격
- 서울지방검찰청 검사
- 법무부 법무실 인권과 검사
- 대전지방검찰청 검사장
- 광주고등검찰청 검사장
- 서울고등검찰청 검사장
- 헌법재판소 재판관
- 고위공직자범죄수사처 자문위원회 위원장
- 법무법인(유한) 화우 고문변호사

[학력사항]

- 서울대학교 사회과학대학 사회학과 졸업
- 서울대학교 대학원 법학과 석사과정 수료

[Career]

- Prosecutor, Seoul District Prosecutors' Office
- Prosecutor, Human Rights Division, Ministry of Justice
- Chief Prosecutor, Daejeon Prosecutors' Office
- Chief Prosecutor, Gwangju High Prosecutors' Office
- Chief Prosecutor, Seoul High Prosecutors' Office
- Justice, Constitutional Court
- Chairperson, Advisory Committee of the Corruption Investigation Office for High-ranking Officials
- Counselor, Hwawoo(Yoon&Yang) Law Firm

[Education]

- B.A. in Sociology, College of Social Science, Seoul National University
- LLM, Seoul National University

개회사

국내외 귀빈 여러분 안녕하십니까,

대한민국 국가인권위원회 위원장 안창호입니다.

대한민국 국가인권위원회와 주한유럽연합대표부, 아시아 태평양 국가인권기구 포럼이 공동으로 주최하는 “신기술과 인권: 인공지능의 기회와 도전” 국제 콘퍼런스에 참석해 주신 여러분들을 진심으로 환영합니다.

오늘 이 자리를 빛내주신,
△ 롤란트 호네캄프 주한유럽연합대표부 부대사님

직접 참석은 어려웠지만, 영상으로 축하의 말씀을 전달해 주신
△ 사마르 하지 하산 아시아 태평양 국가인권기구포럼(APF) 의장님

그리고 기조 연설을 맡아주신,
△ 필립 알스톤 뉴욕대학교 로스쿨 존 노톤 포메로이 교수님,

그리고 각 세션에서 좌장과 발표를 맡아주신 국내외 전문가 여러분께 깊은 감사의 말씀을 드립니다.

존경하는 내외빈 여러분,

우리는 지금 인공지능을 비롯한 신기술이 사회 전반을 빠르게 변화시키는 시대에 살고 있습니다. 인공지능은 의료, 교육, 환경, 노동 등 다양한 분야에서 편익을 제공하는 동시에, 인간의 판단을 대신하며 우리의 권리와 자유에 직접적인 영향을 미치고 있습니다.

그러나 그 영향은 언제나 긍정적인 것만은 아닙니다.

편향된 알고리즘은 차별을 강화하고, 데이터 독점은 불평등을 심화시키며, 무분별한 자동화는 노동권과 고용 안정을 위협하고 있습니다. 이는 단순히 기술적 문제가 아니라, 생명권·평등권·노동권 등 핵심 인권 가치와 직결된 문제입니다.

국제사회는 이러한 도전에 대응하기 시작했습니다.

유엔 인권이사회는 2019년 제41차 회기에서 처음 ‘신기술과 인권’ 결의안을 채택한 이후, 결의를 갱신해 왔으며, 2025년 제59차 회기에는 네 번째 채택이 이루어졌습니다.

유럽연합은 2024년 세계 최초로 AI 법안을 제정하였고, 대한민국 역시 같은 해 AI 기본법을 통과시키며 투명성과 인권 보호를 위한 제도적 기반을 마련했습니다.

우리 국가인권위원회도 2022년「인공지능 인권 가이드라인」과 2024년「AI 인권영향평가 도구」를 마련하였고, AI 법제화 과정에 적극적으로 의견을 제시하며 기술 발전과 인권의 조화를 위해 노력해 왔습니다.

그럼에도 불구하고, 기술 발전은 법과 제도의 속도를 앞질러 가고 있습니다.

따라서 우리는 오늘, 인공지능과 신기술이 가져온 변화, 앞으로 예측되는 미래의 위험과 가능성, 그리고 이에 대응하기 위한 구체적 방안을 함께 논의해야 합니다.

이 점에서 이번 콘퍼런스와 기조연설과 각 세션 발제들은 매우 의미가 큽니다.

필립 알스톤 교수님의 기조연설은 기술 발전과 불평등, 인권 문제의 연결점을 진단할 것이며, 이어지는 세션에서는 국제적 규범과 국가별 대응, 차별과 불평등의 문제, 인권영향평가와 거버넌스 등 실질적 대안을 다루게 될 예정입니다.

이러한 논의는 신기술 시대의 인권 보장 방향을 모색하고, 국제적 협력과 연대를 강화하는 데 중요한 밑거름이 될 것입니다.

존경하는 내외빈 여러분,

오늘의 논의가 신기술 발전에 대한 인권적 대응을 강화하는 계기가 되기를 바랍니다. 기술은 끊임없이 변화하지만, 인권은 변하지 않는 우리의 기준입니다. 오늘 이 만남이 그 기준을 더욱 굳건히 하는 기회가 되기를 기대합니다.

오늘 나눈 지혜가 인권을 위한 구체적 실천으로 이어지기를 소망합니다.

다시 한번 오늘 콘퍼런스에 참석해 주신 여러분, 진심으로 환영과 감사의 말씀을 전합니다.

Opening Remark

Distinguished guests from home and abroad,

I am Chang-ho Ahn, Chairperson of the National Human Rights Commission of Korea (NHRCK).

It is my great honor to welcome you to the International Conference on "New Technologies and Human Rights: Opportunities and Challenges of Artificial Intelligence," co-organized by the NHRCK, the Delegation of the European Union to the Republic of Korea, and the Asia Pacific Forum of National Human Rights Institutions (APF).

I extend my sincere appreciation to those who have graciously joined us today:

Mr. Roland Honekamp, Deputy Head of Mission of the Delegation of the European Union to the Republic of Korea,

Ms. Samar Haj Hassan, Chairperson of the APF, who has shared her congratulatory message via video,

Professor Philip Alston, John Norton Pomeroy Professor of Law at New York University School of Law, who will deliver today's keynote speech,

and all the distinguished experts from Korea and abroad who will serve as chairs and speakers in our sessions.

Ladies and gentlemen,

We are living in a time when new technologies, particularly artificial intelligence, are rapidly reshaping every aspect of our society. AI offers significant benefits across diverse fields—including healthcare, education, the environment, and labor. At the same time, it increasingly substitutes human judgment and exerts direct influence on our rights and freedoms. These impacts, however, are not always positive.

Biased algorithms reinforce discrimination. Data monopolies deepen inequality. Unchecked automation threatens labor rights and job security. These are not merely technical challenges but issues that strike at the heart of fundamental human rights, including the rights to life, equality, and work.

The international community has begun to respond. Since the Human Rights Council first adopted a resolution on "New Technologies and Human Rights" at its 41st session in 2019, the resolution has been renewed, most recently for the fourth time at the 59th session in 2025. The European Union, in 2024, became the first to adopt comprehensive AI legislation. That same year, Korea enacted its Framework Act on AI, establishing an institutional foundation for transparency and the protection of human rights.

The NHRCK has also endeavored to contribute to these efforts. We issued the AI Human Rights Guidelines in 2022, followed by the AI Human Rights Impact Assessment Tool in 2024, and actively provided input throughout the legislative process on AI in Korea, striving to ensure that technological development advances in harmony with human rights.

Nevertheless, the pace of technological innovation continues to outstrip the development of laws and institutions.

It is therefore imperative that we use this conference to reflect together on the profound changes brought about by AI and other new technologies-the risks and opportunities they present-and the concrete measures required in response.

In this regard, today's keynote address and sessions are of great significance. Professor Alston's keynote will examine the links between technological development, inequality, and human rights. The sessions that follow will consider international frameworks and national responses, challenges related to discrimination and inequality, and approaches to human rights impact assessment and governance.

These discussions will play an essential role in shaping the direction of human rights protection in the age of new technologies, while also reinforcing international cooperation and solidarity.

Distinguished guests,

I hope that today's discussion will serve as a turning point in strengthening our human rights–based responses to technological advances. Technologies may evolve without pause,

but human rights remain our enduring standard. May this gathering be an opportunity to reaffirm and fortify that standard.

It is my sincere wish that the wisdom shared here today will be translated into concrete actions for the advancement of human rights.

Once again, I warmly welcome you to this conference and express my deepest gratitude for your participation.

축사 1 | Welcome Remark 1

롤란트 호네캄프

Roland HONEKAMP

주한 EU 대표부 차석대사

Deputy Chief of Mission, Delegation of the EU to the ROK

[주요경력]

롤란트 호네캄프는 2025년 8월부터 주한 유럽연합대표부 차석대사로 임무를 시작했습니다. 그는 2011년 창설된 유럽대외관계청(EEAS)에서 근무를 시작했습니다. 가장 최근에는 브뤼셀에서 일본, 한반도, 호주, 뉴질랜드 및 태평양 도서를 담당하는 부국장을 역임했습니다. 이 직책에서 그는 유럽연합과 주요 파트너 간의 정상회담 및 고위급 회의 준비·지원, G7·G20·UN 협력 업무, 유럽연합의 안보·국방 파트너십 협상, 그리고 북한 관련 업무에 기여했습니다. 그 이전에는 2019년부터 2023년까지 도쿄 주일 유럽연합대표부에서 정치 담당관(Head of Political Section) 을 역임했습니다. 브뤼셀 본부에서는 아시아·태평양 및 중앙아시아 담당 부서와 경제정책 부서에서 다양한 직책을 수행했습니다. 또한 2011년부터 2015년까지는 모스크바 주러시아 유럽연합대표부에서 참사관으로 근무했습니다. 유럽연합 외교부에 합류하기 전, 호네캄프는 브뤼셀의 유럽위원회에서 근무했습니다.

호네캄프는 독일 국적자로, 영국 옥스퍼드대학교 베일리얼 칼리지에서 학사 및 석사 학위를, 런던정치경제대학교(LSE)에서 석사 학위를 취득했습니다. 그의 아내 레카는 유럽위원회에서 근무하는 경쟁법 전문 변호사이며, 두 사람 사이에는 세 명의 10대 자녀가 있습니다.

[Career]

Roland Honekamp started his assignment as Deputy Chief of Mission at the Delegation of the European Union in Seoul in August 2025. He is an official with the External Action Service of the European Union (EEAS), which he joined at its creation in 2011. Mr. Honekamp most recently served as Deputy Head of Division responsible for Japan, the Korean Peninsula, Australia, New Zealand and the Pacific Islands in Brussels. In that capacity, he contributed to the organization of bilateral Summits and leaders meetings with key partners; G7/G20/UN coordination; the negotiation of the EU's security and defence partnerships; and DPRK issues. Previously, he served as Head of Political Section at the Delegation of the European Union to Japan in Tokyo, from 2019 to 2023. In Brussels, he has held various positions in the departments responsible for Asia-Pacific and for Central Asia, as well as in economic policy making. From 2011 to 2015 he was posted to Russia, as counsellor at the Delegation of the European Union in Moscow.

Before joining the foreign service of the European Union in 2011, Mr. Honekamp was an official at the European Commission in Brussels. Mr. Honekamp, a German national, holds a B.A. and M.A. from Balliol College, Oxford University, UK, as well as an M.Sc. from the London School of Economics and Political Science (LSE), UK. He and his wife Reka, an antitrust lawyer with the European Commission, have three teenage children.

축사 2 | Welcome Remark 2

사마르 알 하지 하산

Samar Al Haj HASSAN

아시아 태평양 국가인권기구포럼 의장, 요르단 국가인권기구 위원장
Chairperson, Asia Pacific Forum of National Human Rights Institutions
Chairperson, Jordan National Centre for Human Rights

[주요경력]

하지 하산은 스위스 아메리칸 대학교에서 국제경영과 마케팅을 전공하였습니다. 그녀는 인권을 비롯한 인간·사회·정치 발전 분야에서 30년이 넘는 경력을 가지고 있습니다. 하지 하산의 경험은 공공 및 민간 부문뿐 아니라 국제적·국내적 영역을 아우르며, 특히 가족·청년·여성을 위한 전략적·운영 계획의 수립과 발전에 기여해왔습니다. 요르단 하심 왕국이 건국 2세기를 맞이하는 시점에, 하지 하산은 자신의 지식과 실무 경험을 바탕으로 정치체제 현대화를 위한 왕립 위원회에 참여하였으며, 여성 역량강화위원회 위원장으로 선출되었습니다. 하지 하산은 UNIFEM, 라니아 알 압둘라 요르단 왕비 사무국, USAID-요르단, Vital Voices, ECODIT, UNICEF, 요르단 리버 재단, 사회개발부, 교육부, 세계은행, UNFPA 등 여러 국내외 기관에서 자문 역할을 맡아왔습니다.
그녀는 여성, 청년, 아동, 그리고 전반적인 가족 복지 증진을 위해 헌신적으로 활동해 왔습니다. 2009년 요르단 사회경제위원회 위원으로 임명되었으며, 같은 해부터 현재까지 요르단 국제여성포럼 회원으로 활동하고 있습니다. 2011년 10월에는 요르단 의회 상원의원으로 임명되었고, 2014년부터 2020년까지는 독립선거위원회 위원으로 활동했습니다. 2020년에는 유럽선거지원센터 자문위원회 위원으로 위촉되었으며, 2022년 10월에는 요르단 국가인권센터 의장으로 임명되었습니다. 2023년 8월에는 이사회 의장으로 재임명되어, 4년 임기를 이어나가고 있습니다.

[Career]

Mrs. Haj Hassan specialized in international business and marketing from the American University of Switzerland. Her experience in the field of human, social, and political development and human rights extends over 30 years. Mrs. Haj Hassan's experiences vary between the public and private sectors, and in international and national frameworks. She has contributed primarily to the design and development of strategic and operational plans for the family, youth, and women. In conjunction with the entry of the Hashemite Kingdom of Jordan into its second centenary. Mrs. Haj Hassan contributed to the translation of her knowledge and practical experiences into her membership of the Royal Committee for the Modernization of the Political System, where she was elected as Chairperson of the Committee on the Empowerment of Women. Mrs. Haj Hassan served as an adviser to several national and international organizations, including UNIFEM, Her Majesty Queen Rania Al-Abdullah's Office, USAID-Jordan, Vital Voices, ECODIT, UNICEF, Jordan River Foundation, Ministry of Social Development, Ministry of Education, World Bank and UNFPA. Mrs. Haj Hassan is a strong advocate for women, youth, children, and the welfare of the family in general. In 2009, she was appointed a member of the Social and Economic Council in Jordan, and since 2009 till present she joined as a member of the International Women Forum in Jordan. In October 2011, she was appointed as a member of the upper senate in the Jordanian Parliament. Between 2014-2020, she was appointed as a Commissioner in the Independent Election Commission. In 2020 she became a member of the Committee of Advisers of the European Centre for Electoral Support, and in October 2022 she was appointed a Chairperson of the National Center for Human Rights and her appointed was renewed in August 2023 as Chairperson of the Board of Trustees for four years.

축사 2

여러분, 안녕하십니까.

오늘 이 자리에 함께할 수 있어 매우 기쁘게 생각합니다. 이번 회의를 주최해 주신 대한민국 국가인권위원회와 모든 공동 주최자 여러분께 깊이 감사드립니다.

오늘날 AI가 가져온 급격한 변화는 우리가 경험한 적 없는 새로운 인권의 기회와 도전을 동시에 제시하고 있습니다.

우리는 AI가 인류에 막대한 이익을 가져다줄 잠재력을 지니고 있음을 잘 알고 있습니다. AI는 전략적 예측과 전망 능력을 향상시키고, 지식 접근을 민주화하며, 과학적 진보를 가속화하고, 방대한 정보 처리 역량을 증대시킬 수 있습니다. AI는 특히 보건과 교육과 같은 분야에서 인권을 증진하는 강력한 도구가 될 수 있지만, 동시에 사생활 침해, 표현의 자유 위축, 차별로 이어질 수 있는 잠재적 편향에 대한 우려도 있습니다. 따라서 AI 기술이 인권을 존중하는 방식으로 개발·활용되도록 하기 위해서는 규제적·비규제적 조치를 포함한 효과적인 AI 거버넌스가 필수적입니다.

예를 들어, AI의 기회의 측면에서는 다음과 같은 것들을 이야기할 수 있습니다.

- AI는 인권 침해를 모니터링하고, 사법 접근성을 개선하며, 보건의료와 교육을 강화할 수 있습니다.
- AI는 장애인 권리를 지원하고, 정보 접근을 확대하며, 자립적인 생활을 가능하게 할 수 있습니다.
- AI는 정부 활동을 감시하고, 투명성을 제고하며, 공공 서비스를 개선할 수 있습니다.

반면 AI의 도전의 측면에서는, 아래와 같은 것들이 이야기됩니다.

- 기존의 편견을 강화·확대하여 특히 소외된 집단에 불공정하거나 차별적인 결과를 초래할 수 있습니다.
- 안면인식 및 데이터 분석 기술은 감시와 통제에 악용될 수 있어 사생활과 표현의 자유를 침해할 수 있습니다.
- 소셜미디어에서의 AI 기반 콘텐츠 조정은 합법적인 표현을 억압하거나 편향된 담론 공간을 형성할 수 있습니다.
- AI는 딥페이크와 허위정보를 생성하여 공공 신뢰와 민주적 절차를 훼손할 수 있습니다.

- AI에 과도하게 의존할 경우 오류가 발생하고, 인간의 판단력이 약화될 수 있습니다.

이러한 기회를 활용하고 도전 과제에 대응하기 위해서 우리는 아래와 같은 사안들을 유념해야 합니다.

- AI 개발과 활용 전 과정에서 인권을 최우선으로 하는 규제와 제도를 마련해야 합니다.
- AI 시스템이 투명하고 설명 가능하며, 그 결과에 대해 책임을 질 수 있도록 보장해야 합니다.
- AI 정책 수립에 있어 시민사회, 인권단체, 그리고 소외된 집단의 참여와 협의를 보장해야 합니다.
- AI 시스템 편향을 최소화 할 수 있도록 다양한 배경을 지닌 팀에 의해 개발되고, 다양한 데이터를 기반으로 학습되도록 해야 합니다.
- 인간의 안녕을 최우선으로 하고 인권을 존중하는 윤리적 AI 개발 문화를 조성해야 합니다.

이러한 도전을 극복하고 기회를 적극적으로 활용한다면, 우리는 AI의 힘을 통해 인권을 증진하고 더욱 정의롭고 공정한 사회를 만들어갈 수 있을 것입니다.

인권 원칙이 확고히 자리 잡는다면, 우리는 AI의 놀라운 가능성을 안전하게 활용하며, 인간이 발휘할 수 있는 창의성과 혁신, 그리고 최고의 성과를 지속적으로 추구할 수 있을 것이라 확신합니다.

감사합니다.

Welcome Remark 2

Good morning,

It is a pleasure to speak with you today. My warm thanks to the National Human Rights Commission of Korea and all of the co-sponsors who are hosting this conference.

Today the rapid and seismic shifts caused by Artificial Intelligence (AI) are presenting opportunities and challenges for human rights that none of us have ever encountered before.

We know that AI has the potential to be enormously beneficial to humanity. It could improve strategic foresight and forecasting, democratize access to knowledge, turbocharge scientific progress, and increase capacity for processing vast amounts of information. While AI can be a powerful tool for advancing human rights, particularly in areas like healthcare and education, it also raises concerns about privacy, freedom of expression, and potential biases that could lead to discrimination. Effective AI governance, including regulatory and non-regulatory measures is, therefore, crucial to ensure AI technologies are developed and deployed in a way that respects human rights.

For example, in terms of the opportunities of AI:

- AI can be used to monitor human rights violations, improve access to justice, and enhance healthcare and education.
- AI can assist people with disabilities, provide access to information, and support independent living.
- AI-powered tools can be used to monitor government actions, promote transparency, and improve public services.

And in terms of the challenges of AI:

- AI systems can perpetuate and amplify existing biases, leading to unfair or discriminatory outcomes, particularly for marginalized groups.
- AI technologies like facial recognition and data analytics can be used for surveillance and

monitoring, raising concerns about privacy and freedom of expression.

- AI-driven content moderation on social media platforms can inadvertently suppress legitimate forms of expression or create echo chambers.
- AI can be used to generate deepfakes and other forms of disinformation, potentially undermining public trust and democratic processes.
- Over-reliance on AI-informed decisions can lead to errors and undermine human judgment.

To address the challenges and harness the opportunities, we need to:

- Implement regulations and frameworks that prioritize human rights in the development and deployment of AI.
- Ensure that AI systems are transparent, explainable, and accountable for their decisions.
- Engage with civil society, human rights groups, and marginalized communities in the development of AI policies.
- Ensure that AI systems are developed by diverse teams and trained on diverse datasets to mitigate bias.
- Foster a culture of ethical AI development that prioritizes human well-being and respects human rights.

By addressing these challenges and seizing the opportunities, societies can harness the power of AI to advance human rights and create a more just and equitable world.

With human rights principles firmly in place, I believe we can safely harness AI's incredible opportunities and keep striving for creativity, innovation and the best that human beings can deliver.

Thank you.

신기술과 인권: 인공지능의 기회와 도전

New Technology and Human Rights: Opportunities and Challenges of Artificial Intelligence

기조연설

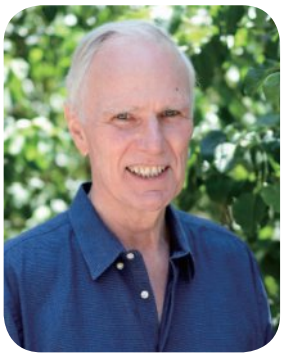
Keynote Speech

발표자
Speaker

필립 알스톤 | 뉴욕대학교 로스쿨 존 노톤 포메로이 교수, 전 유엔 빈곤과 인권 특별보고관

Philip ALSTON | John Norton Pomeroy Professor of Law, School of Law, New York University
Former UN Special Rapporteur on Extreme Poverty and Human Rights

기조연설 | Keynote Speech



필립 알스톤

Philip ALSTON

뉴욕대학교 로스쿨 존 노톤 포메로이 교수, 전 유엔 빈곤과 인권 특별보고관
John Norton Pomeroy Professor of Law, School of Law, New York University
Former UN Special Rapporteur on Extreme Poverty and Human Rights

[주요경력]

필립 알스톤은 뉴욕대학교 법학대학원의 존 노턴 포메로이 교수이다. 그는 UN 빈곤과 인권 특별보고관(2014-2020), UN 비사법적 처형 특별보고관(2004-2010), UN 경제적·사회적·문화적 권리위원회 위원장(1991-1998), 중앙아프리카공화국에 관한 안전보장이사회 조사위원회위원(2014-2015), 아동권리협약 초안 작성 시 유니세프 법률고문(1986-1990), UN 인권 조약기구 개혁 독립전문가(1989-1997), 새천년개발목표에 관한 UN 인권최고대표사무소 특별고문(2004-2007)을 역임했다. 그는 https://humanrightstextbook.org/에서 무료로 이용할 수 있는 포괄적 교과서인 『국제인권』의 저자이다.

[Career]

Philip Alston is John Norton Pomeroy Professor at New York University Law School. He was UN Special Rapporteur on extreme poverty and human rights (2014-2020), UN Special Rapporteur on extrajudicial executions (2004-2010), Chairperson, UN Committee on Economic, Social and Cultural Rights (1991-1998), Member of the Security Council Commission of Inquiry on the Central African Republic (2014-2015), legal adviser to UNICEF in drafting the Convention on the Rights of the Child (1986-1990), Independent Expert on reform of the UN human rights treaty body system (1989-1997), and Special Advisor to the UN High Commissioner for Human Rights on the Millennium Development Goals (2004-2007).

He is the author of International Human Rights, a comprehensive textbook available free online at https://humanrightstextbook.org/

[기조연설 | Keynote Speech]

AI 발전 과정에 인권을 반영하기

Making Human Rights Relevant in Shaping AI

필립 알스톤

Philip ALSTON

뉴욕대학교 로스쿨 존 노톤 포메로이 교수, 전 유엔 빈곤과 인권 특별보고관

John Norton Pomeroy Professor of Law, School of Law, New York University

Former UN Special Rapporteur on Extreme Poverty and Human Rights

AI 발전 과정에 인권을 반영하기

필립 알스톤 | 뉴욕대학교 로스쿨 존 노톤 포메로이 교수, 전 유엔 빈곤과 인권 특별보고관

1. 인공지능 정책과 인권의 현 주소

- 거대 기술 기업- 실리콘 밸리 기준의 좁은 관점
- 중국- '인공지능+' 정책 추진을 위한 국무원 의견 발표 (2025년 8월 21일)
- 미국- ' 인공지능 행동계획 :경쟁에서 승리하기' 발표 (2025년 7월)
- 유럽연합- 2024년 EU 인공지능법 통과, 2026년 8월 2일까지 인공지능 거버넌스 시스템 구축 예정
- 기타

2. 인권계의 반응

- 유엔 사무총장 – 2024년 글로벌 디지털 콤팩트
- 유엔인권최고대표 – 정책 제안, 인권최고대표사무소의 역할
- 그 외 유엔 전문가 집단 – 특별절차, 조약기구
- 시민단체 – 주요 국제 비정부 기구
- 우려사항과 해결방안 개요

3. 현 대응방식의 한계

- 여러 노력들이 단절적, 분절적으로 이루어지는 상황
- 외부 중심이 아니라 내부 중심의 관점
- 기술적 측면의 경로의존성
- 모니터링 의지 낮음

4. 발전 방향

- 인공지능은 기존의 방식과 다름
- 더욱 방대한 제도적 틀이 필요
  - 인간의 존엄성 – 포괄적 개념이 필요함
  - 개인 vs 집단 – 규범적 개인주의를 초월해야 함
  - 경제적, 사회적 권리 – 소외의 비용
  - 개발권 – 인공지능의 맥락에서 유용할 것인지 고민해야

4. 발전 방향 (2)

- '파트너십' 이상이 필요함
  - 유엔 기업과 인권 이행원칙(UNGP)에 대한 현실주의적 태도 - 기업과 인권 실무그룹
- 구조적 접근
  - 불평등, 재정정책

Making Human Rights Relevant in Shaping AI

Philip ALSTON | John Norton Pomeroy Professor of Law, School of Law, New York University
Former UN Special Rapporteur on Extreme Poverty and Human Rights

1. Current place of human rights in AI policies

- Big Tech – the tunnel vision of Silicon Valley
- China – Opinion of the State Council on the In-depth Implementation of the 'Artificial Intelligence +' Action, 21 August 2025
- United States – Winning the Race: America's AI Action Plan, July 2025
- European Union – The EU AI Act of 2024; Establishing an AI governance system by 2 August 2026
- The rest

2. Responses by the human rights community

- The UN Secretary-General – Global Digital Compact, 2024
- The UN High Commissioner for Human Rights – Policy proposals, Role of the Office
- U.N. experts – Special procedures, Treaty bodies
- Civil society – Leading INGOs
- Overview of concerns and solutions

3. Shortcomings in the responses

- Siloing and dispersion of efforts
- Inward rather than outward looking
- Path dependency in terms of techniques
- Reticence in monitoring

4. Ways forward

- AI is different – Business-as-usual won't work
- A broader normative framework
  - Human dignity – the need for an umbrella concept
  - Individual versus collective – Transcending normative individualism
  - Economic and social rights – The costs of marginalization
  - The right to development – Could it be useful in this context?

4. Ways forward (2)

- Beyond 'partnership'
  - Realism about the UNGPs – The Working Group on Business and Human Rights
- A structural approach
  - Inequality, Fiscal policy

세션 1
Session 1

신기술과 인권: 국제 동향과 국가별 대응
New Technology and Human Rights: Global Trends and National Responses

사회자 Moderator

키렌 피츠페트릭 | 아시아 태평양 국가인권기구포럼 사무국장
Kieren FITZPATRICK | Director, Asia Pacific Forum of National Human Rights Institutions

발표자 Speaker

백범석 | 경희대학교 법학전문대학원 교수, 유엔 인권이사회 자문위원회 위원
BAEK Buhm-suk | Professor, School of Law, Kyunghee University
Member, UN Human Rights Council Advisory Committee

로베르토 지커리 | 서울대학교 데이터사이언스대학원 특임교수
Roberto V. ZICARI | Adjunct Professor, Graduate School of Data Science, Seoul National University

안진현 | 국가인권위원회 인권정책과 사무관
AHN Jin-hyun | Deputy Director, Human Rights Policy Division, NHRCK

세션 1 | Session 1

[사회자_Moderator]

키렌 피츠페트릭

Kieren FITZPATRICK

아시아 태평양 국가인권기구포럼 사무국장

Director, Asia Pacific Forum of National Human Rights Institutions

[주요경력]

키렌 피츠페트릭 이사는 사회과학, 철학, 법학 학위를 보유하고 있으며, 아시아 태평양 국가인권기구포럼의 창립 이사다.

[Career]

Mr Kieren Fitzpatrick has degrees in social science, philosophy and law. He is the foundation Director of the Asia Pacific Forum of National Human Rights Institutions.

세션 1 | Session 1

[발표자_Speaker]

백범석

BAEK Buhm-suk

경희대학교 법학전문대학원 교수, 유엔 인권이사회 자문위원회 위원
Professor, School of Law, Kyunghee University
Member, UN Human Rights Council Advisory Committee

[주요경력]

경희대학교 법학전문대학원 교수로 재직 중이며, 현재 유엔 인권이사회 자문위원회 위원 및 대법원 양형위원회 위원으로 활동하고 있다. 국제공법 및 국제인권법에 관한 30여편의 연구논문과 저서를 출간하였고, 2021년 보고관으로서 디지털 신기술과 인권에 관한 연구보고서(A/HRC/47/52)를 유엔 인권이사회에 제출하였고, 현재는 인공지능이 굿거버넌스에 미치는 영향에 관한 보고서를 준비 중에 있다(A/HRC/RES/57/5). 서울대학교에서 법학학사(LL.B.), 연세대학교에서 국제학석사(M.A.), 코넬대학교에서 법학석사(LL.M.), 법학박사(J.S.D.)를 취득하였다.

[Career]

Buhm-Suk BAEK is a Professor at Kyung Hee University Law School in Korea. His research focuses on International Human Rights Law and Public International Law. He has published over 30 research papers and books, including "Transnational Justice in Unified Korea" with Ruti Teitel. Currently, he serves as a member of the Advisory Committee of the UN Human Rights Council and as a member of the Sentencing Commission of the Supreme Court of Korea. In 2021, as Rapporteur, he submitted the report "Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights" (A/HRC/47/52) to the UN Human Rights Council. He is currently preparing a report on the impact of artificial intelligence on good governance (A/HRC/RES/57/5). He has served as an advisor to various government agencies including the National Human Rights Commission of Korea, Ministry of Justice, Ministry of Unification, and Ministry of National Defense. He has also served as a standing director of the Korean Society of International Law and research director of the Korean Association of Human Rights Law. He received an LL.B. from Seoul National University, an M.A. in International Studies from Yonsei University, and an LL.M. and J.S.D. from Cornell Law School. (205 words)

세션 1 | Session 1

[발표자_Speaker]

로베르토 지커리

Roberto V. ZICARI

서울대학교 데이터사이언스대학원 특임교수

Adjunct Professor, Graduate School of Data Science, Seoul National University

[주요경력]

로베르토 V. 지커리 교수는 핀란드 헬싱키의 아르카다 전문대학(Yrkeshögskolan Arcada)에서 객원 교수로, 한국 서울대학교 데이터사이언스대학원에서 특임교수로 활동하고 있습니다.

그는 또한 유럽대학교연구소(European University Institute) 로버트 슈만 고등연구센터(Robert Schuman Centre for Advanced Studies)의 방문 펠로우로 활동하고 있습니다.

지커리 교수는 국제 전문가 팀을 이끌며 '신뢰할 수 있는 인공지능(Trustworthy AI)' 평가 프로세스인 Z-Inspection®을 마련했습니다.

과거에는 독일 프랑크푸르트 괴테대학교에서 데이터베이스 및 정보시스템(DBIS) 교수로 재직하며 프랑크푸르트 빅데이터 연구소(Frankfurt Big Data Lab)를 설립했습니다.

그는 데이터베이스와 빅데이터 분야에서 국제적으로 인정받는 전문가이며, 윤리와 인공지능, 혁신, 기업가 정신에도 폭넓은 관심을 가지고 있습니다. 또한 ODBMS.org 웹 포털과 ODBMS Industry Watch 블로그의 편집장을 맡고 있습니다. 몇 년 동안 미국 UC 버클리(캘리포니아대학교 버클리) 산업공학 및 운영연구학과 산하 기업가정신·기술센터(Center for Entrepreneurship and Technology)에서 객원 교수로 활동한 바 있습니다.

[Career]

Professor Roberto V. Zicari is an affiliated professor at the Yrkeshögskolan Arcada, Helsinki, Finland, and an adjunct professor at the Seoul National University, South Korea.

He is also Visiting Fellow at the Robert Schuman Centre for Advanced Studies at the European University Institute.

Roberto V. Zicari is leading a team of international experts who defined an assessment process for Trustworthy AI, called Z-Inspection®.

Previously he was professor of Database and Information Systems (DBIS) at the Goethe University Frankfurt, Germany, where he founded the Frankfurt Big Data Lab .

He is an internationally recognized expert in the field of Databases and Big Data. His interests also expand to Ethics and AI, Innovation and Entrepreneurship. He is the editor of the ODBMS.org web portal and of the ODBMS Industry Watch Blog. He was for several years a visiting professor with the Center for Entrepreneurship and Technology within the Department of Industrial Engineering and Operations Research at UC Berkeley (USA).

세션 1 | Session 1

[발표자_Speaker]

안진현

AHN Jin-hyun

국가인권위원회 인권정책과 사무관

Deputy Director, Human Rights Policy Division, NHRCK

[주요경력]

2005년부터 국가인권위원회에서 정보화, 인권침해 및 차별 진정사건 조사, 인권교육 등 업무를 수행해왔습니다. 현재는 인공지능 등 신기술, 개인정보 보호, 통신의 비밀 보호 등 정보인권 관련 정책과 법제를 검토하는 업무를 담당하고 있습니다.

[Career]

Since 2005, Jin-hyun AHN has been working at the National Human Rights Commission of Korea, where she has handled various tasks, including digitalization efforts, investigating human rights violations and discrimination cases, and providing human rights education. Currently, she is responsible for reviewing policies and legal systems related to digital and human rights, with a focus on new technologies like artificial intelligence, personal data protection, and the protection of communication privacy.

[발표 1 | Speaker 1]

신기술과 인권: 글로벌 대응과 인권과제

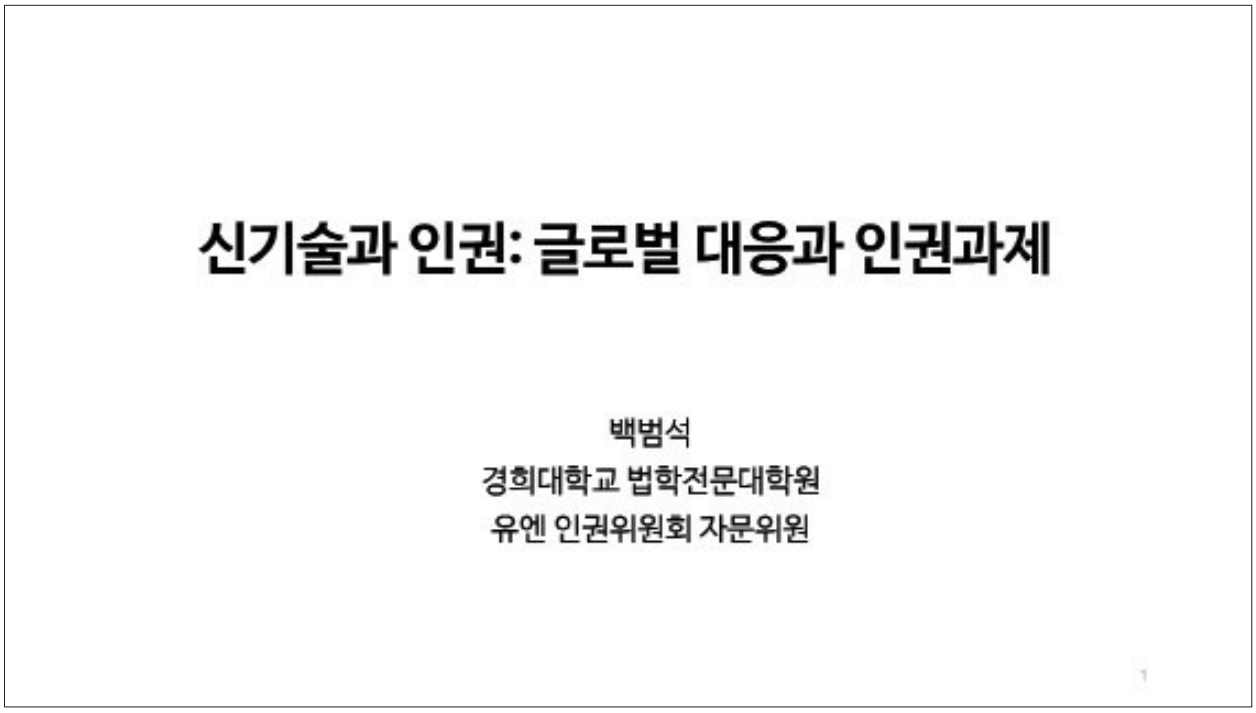
New and Emerging Technologies and Human Rights: Global Responses and Challenges

백범석
BAEK Buhm-suk

경희대학교 법학전문대학원 교수, 유엔 인권이사회 자문위원회 위원
Professor, School of Law, Kyunghee University
Member, UN Human Rights Council Advisory Committee

신기술과 인권: 글로벌 대응과 인권과제

백범석 | 경희대학교 법학전문대학원 교수, 유엔 인권이사회 자문위원회 위원



• 유엔 군축연구소(UNIDIR) 사이버 정책 포털(https://cyberpolicyportal.org/)에 따르면, 2025년 8월 기준 현재까지 아프리카 연합(55개 국가), 에스토니아, 캐나다, 대한민국을 포함하여 30개 이상의 국가 입장 문서(national position paper)가 발표되었으며, 이 가운데 25개 국가는 사이버 공간에서의 국제 인권법 적용 문제를 논의하며, 디지털 기술과 인권의 조화를 이를 필요성을 강조하였다.

*"*인권 존중은 국제 인권법상 확립된 의무이며,이 의무는 사이버 공간에서도 동일하게 적용됩니다. 대한민국은 국제 인권법이 온라인에서도 오프라인과 동일하게 적용됨을 확인합니다. 개인은 다른 모든 영역과 마찬가지로 사이버 활동과 관련해서도 동일한 인권을 누릴 권리가 있습니다. 사생활보호, 표현의 자유, 정보 접근권, 차별로부터의 보호, 혐오 표현 금지 등 기본권은 여성과 사회적 약자를 포함한 모든 개인에게 사이버 공간에서도 보장되고 존중되어야 합니다."*

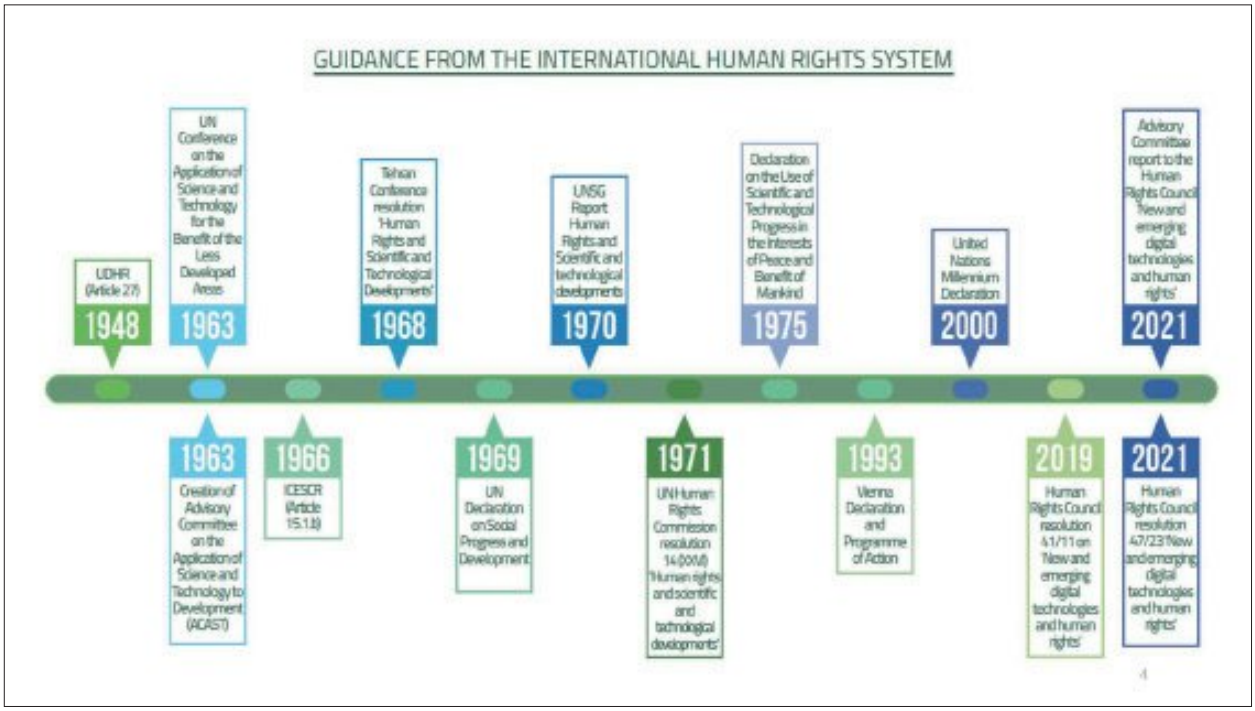
사이버 공간에서의 국제법 적용에 관한 대한민국의 국가 입장(2025)

2

• 글로벌 디지털 콤팩트 (2024)

8(C). 본 콤팩트는 국제 인권법을 포함한 국제법에 근거한다. 시민적·정치적 권리, 경제적·사회적·문화적 권리와 기본적 자유를 포함한 모든 인권은 온라인과 오프라인에서 존중·보호·증진되어야 한다. 우리의 협력은 아동의 권리, 장애인의 권리, 발전권 등을 포함한 모든 인권을 증진하기 위해 디지털 기술을 활용할 것이다;

3



• **세계인권선언 제27조**

제27조

1. 모든 사람은 공동체의 문화생활에 자유롭게 참여하며 예술을 향유하고 과학의 발전과 그 혜택을 공유할 권리를 가진다.

2. 모든 사람은 자신이 창작한 과학적 , 문학적 또는 예술적 산물로부터 발생하는 정신적, 물질적 이익을 보호받을 권리를 가진다.

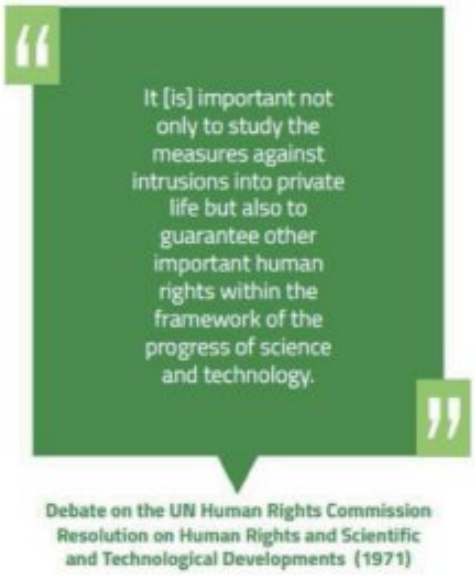
• **경제적, 사회적 및 문화적 권리에 관한 국제 규약 (ICESCR)**

제15조

1. 이 규약의 당사국은 모든 사람의 다음 권리를 인정한다.

(b) 과학의 진보 및 응용으로부터 이익을 향유할 권리

5



인권과 기술의 관계에 대한 논쟁은 현대 국제 인권법이 등장한 이래 오랫동안 이어져 왔다. 신기술의 **양면성** (기회와 도전) 을 다루는 논의 역시 결코 새로운 담론이 아니다. 디지털 기술 자체가 새로운 것이 아니지만, 수년에 걸친 성능의 기하급수적 향상으로, 그 어느 때보다 빠른 속도로 대규모의 복잡한 작업을 수행할 수 있게 되었다. 방대한 데이터의 가용성과 인터넷 및 초고속 통신망의 성장은 인터넷과 광대역 네트워크를 기반으로 한 디지털 신기술의 등장과 활용으로 이어졌으며, 이는 흔히 '제4차 산업혁명'이라고 불린다.

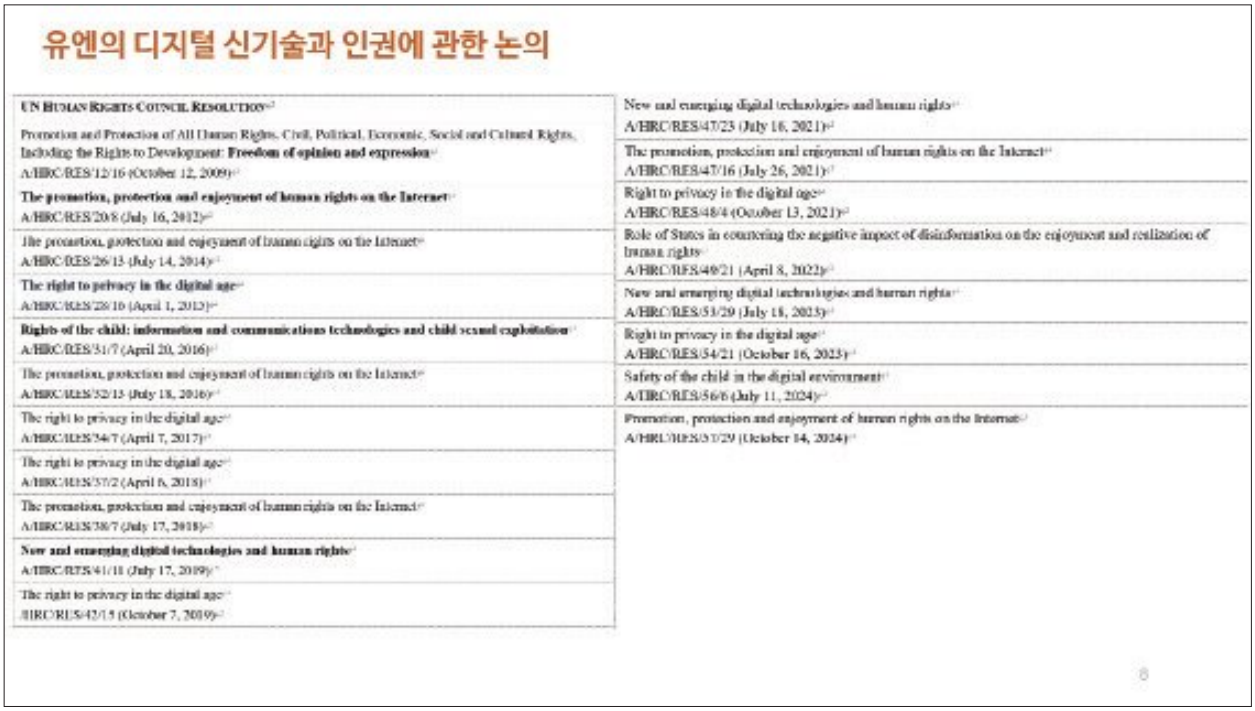
6

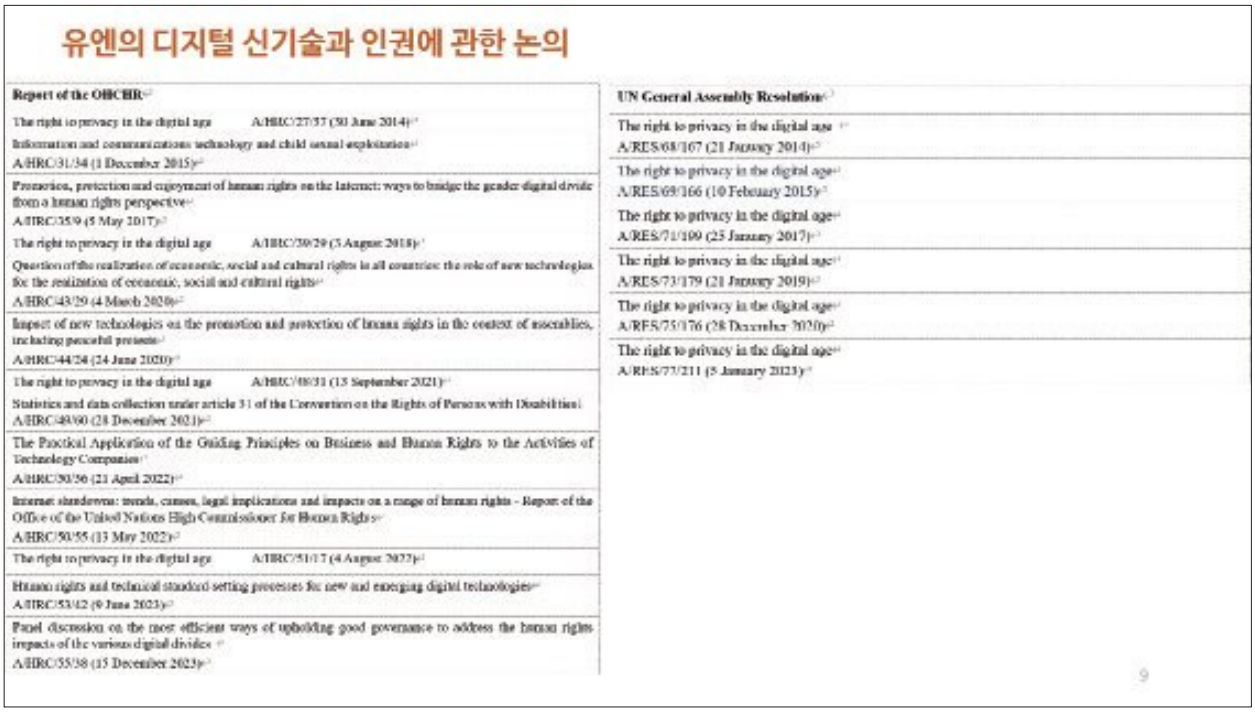
유엔의 디지털 신기술과 인권에 관한 논의

- 지난 20년간 유엔 중심의 국제 인권 메커니즘 체계 하에서 디지털 신기술을 주제로 한 다양한 결의안이 채택되었고, 특별 절차 보고서가 제출되었으며, 조약 기구들의 일반논평 및 권고도 발표되었다. 이러한 논의는 특히 프라이버시권, 집회 및 결사의 자유, 표현의 자유, 아동권, 사회권 등 특정 권리에 미치는 영향을 중심으로 전개되었다.

- 특히 유엔 인권이사회는 2019년 신기술과 인권에 관한 결의안(A/HRC/RES/41/11)을 최초로 채택하였으며, 이 결의안에 따라 자문위원회는 2021년 "인권 증진 및 보호와 관련한 새로운 디지털 기술의 잠재적 영향, 기회 및 도전 과제"에 관한 보고서를 제출하였다.

7


유엔의 디지털 신기술과 인권에 관한 논의
UN HUMAN RIGHTS COUNCIL RESOLUTION
Promotion and Protection of All Human Rights, Civil, Political, Economic, Social and Cultural Rights, Including the Right to Development: Freedom of opinion and expression
A/HRC/RES/12/16 (October 12, 2009)
The promotion, protection and enjoyment of human rights on the Internet
A/HRC/RES/20/8 (July 16, 2012)
The promotion, protection and enjoyment of human rights on the Internet
A/HRC/RES/26/13 (July 14, 2014)
The right to privacy in the digital age
A/HRC/RES/28/16 (April 1, 2015)
Rights of the child: information and communications technologies and child sexual exploitation
A/HRC/RES/31/7 (April 20, 2016)
The promotion, protection and enjoyment of human rights on the Internet
A/HRC/RES/32/13 (July 18, 2016)
The right to privacy in the digital age
A/HRC/RES/34/7 (April 7, 2017)
The right to privacy in the digital age
A/HRC/RES/37/2 (April 6, 2018)
The promotion, protection and enjoyment of human rights on the Internet
A/HRC/RES/38/7 (July 17, 2018)
New and emerging digital technologies and human rights
A/HRC/RES/41/11 (July 17, 2019)
The right to privacy in the digital age
HRC/RES/42/15 (October 7, 2019)
New and emerging digital technologies and human rights
A/HRC/RES/47/23 (July 16, 2021)
The promotion, protection and enjoyment of human rights on the Internet
A/HRC/RES/47/16 (July 26, 2021)
Right to privacy in the digital age
A/HRC/RES/48/4 (October 13, 2021)
Role of States in countering the negative impact of disinformation on the enjoyment and realization of human rights
A/HRC/RES/49/21 (April 8, 2022)
New and emerging digital technologies and human rights
A/HRC/RES/51/29 (July 18, 2023)
Right to privacy in the digital age
A/HRC/RES/54/21 (October 16, 2023)
Safety of the child in the digital environment
A/HRC/RES/56/6 (July 11, 2024)
Promotion, protection and enjoyment of human rights on the Internet
A/HRC/RES/57/29 (October 14, 2024)
8


유엔의 디지털 신기술과 인권에 관한 논의
Report of the OHCHR
The right to privacy in the digital age A/HRC/27/37 (30 June 2014)
Information and communications technology and child sexual exploitation
A/HRC/31/34 (1 December 2015)
Promotion, protection and enjoyment of human rights on the Internet: ways to bridge the gender digital divide from a human rights perspective
A/HRC/35/9 (5 May 2017)
The right to privacy in the digital age A/HRC/39/29 (3 August 2018)
Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights
A/HRC/43/29 (4 March 2020)
Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests
A/HRC/44/24 (24 June 2020)
The right to privacy in the digital age A/HRC/48/31 (13 September 2021)
Statistics and data collection under article 31 of the Convention on the Rights of Persons with Disabilities
A/HRC/49/60 (28 December 2021)
The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of Technology Companies
A/HRC/50/56 (23 April 2022)
Internet shutdowns: trends, causes, legal implications and impacts on a range of human rights - Report of the Office of the United Nations High Commissioner for Human Rights
A/HRC/50/55 (13 May 2022)
The right to privacy in the digital age A/HRC/51/17 (4 August 2022)
Human rights and technical standard-setting processes for new and emerging digital technologies
A/HRC/53/42 (9 June 2023)
Panel discussion on the most efficient ways of upholding good governance to address the human rights impacts of the various digital divides
A/HRC/55/38 (15 December 2023)
UN General Assembly Resolution
The right to privacy in the digital age
A/RES/68/167 (21 January 2014)
The right to privacy in the digital age
A/RES/69/166 (10 February 2015)
The right to privacy in the digital age
A/RES/71/199 (25 January 2017)
The right to privacy in the digital age
A/RES/73/179 (21 January 2019)
The right to privacy in the digital age
A/RES/75/176 (28 December 2020)
The right to privacy in the digital age
A/RES/77/211 (5 January 2023)
9

유엔의 디지털 신기술과 인권에 관한 논의

Report of UN HRC Special Procedures i.e.

Freedom of expression, States and the private sector in the digital age & Private sector roles and public/private regulation (digital spaces)
- Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/32/38 (11 May 2016)

Ending the torment: tackling bullying from the schoolyard to cyberspace
- Special Representative of the Secretary-General on Violence against Children (September 2016)

Right to freedom of opinion and expression
- Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/38/35 (6 April 2018)

Online violence against women and girls from a human rights perspective
- Special Rapporteur on violence against women and girls, its causes and consequences
A/HRC/38/47 (18 June 2018)

Big Data and Open Data
- Special Rapporteur on the right to privacy
A/73/438 (17 October 2018)

Security and surveillance
- Special Rapporteur on the right to privacy
A/HRC/37/62 (25 October 2018)

Data collection and management as a means to create heightened awareness of violence and discrimination based on sexual orientation and gender identity
- Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity
A/HRC/41/45 (14 May 2019)

Right to freedom of peaceful assembly and of association
- Special Rapporteur on the right to freedom of peaceful assembly and of association
A/HRC/41/41 (17 May 2019)

Survelance and human rights
- Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/41/35 (28 May 2019)

Online hate speech
- Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/74/486 (9 October 2019)

Privacy, technology and other human rights from a gender perspective
- Special Rapporteur on the right to privacy
A/HRC/40/63 (16 October 2019)

Security and surveillance, health data, and business enterprises use of personal data
- Special Rapporteur on the right to privacy
A/HRC/43/52 (24 March 2020)

Preliminary evaluation of the privacy dimensions of the coronavirus disease (COVID-19) pandemic
- Special Rapporteur on the right to privacy
A/75/147 (27 July 2020)

Artificial intelligence and privacy, and children's privacy
- Special Rapporteur on the right to privacy
A/HRC/46/37 (25 January 2021)

General Comment No 25(2021) on children's rights in relation to the digital environment
- Committee on the Rights of the Child
CRC/C/GC/25 (2 March 2021)

Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights
- Human Rights Council Advisory Committee
A/HRC/47/52 (19 May 2021)

Disinformation and freedom of opinion and expression
- Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/47/25 (13 April 2021)

How pandemic can be managed with respect to the right to privacy
- Special Rapporteur on the right to privacy
A/76/220 (24 July 2021)

Privacy and personal data protection in Ibero-America: A step towards globalization?
- Special Rapporteur on the right to privacy
A/HRC/49/55 (13 January 2021)

Reinforcing media freedom and the safety of journalists in the digital age
- Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/50/29 (20 April 2022)

Privacy and data protection: Increasingly precious asset in digital era
- Special Rapporteur on the right to privacy
A/77/196 (20 July 2022)

Implementation of the principles of purpose limitation, deletion of data and demonstrate or proactive accountability in the processing of personal data collected by public entities in the context of the COVID-19 pandemic
- Special Rapporteur on the right to privacy
A/HRC/52/37 (27 December 2022)

Principles of transparency and explainability in the processing of personal data in artificial intelligence
- Special Rapporteur on the right to privacy
A/78/310 (30 August 2023)

Legal safeguards for personal data protection and privacy in the digital age
- Special Rapporteur on the right to privacy
A/HRC/55/46 (18 January 2024)

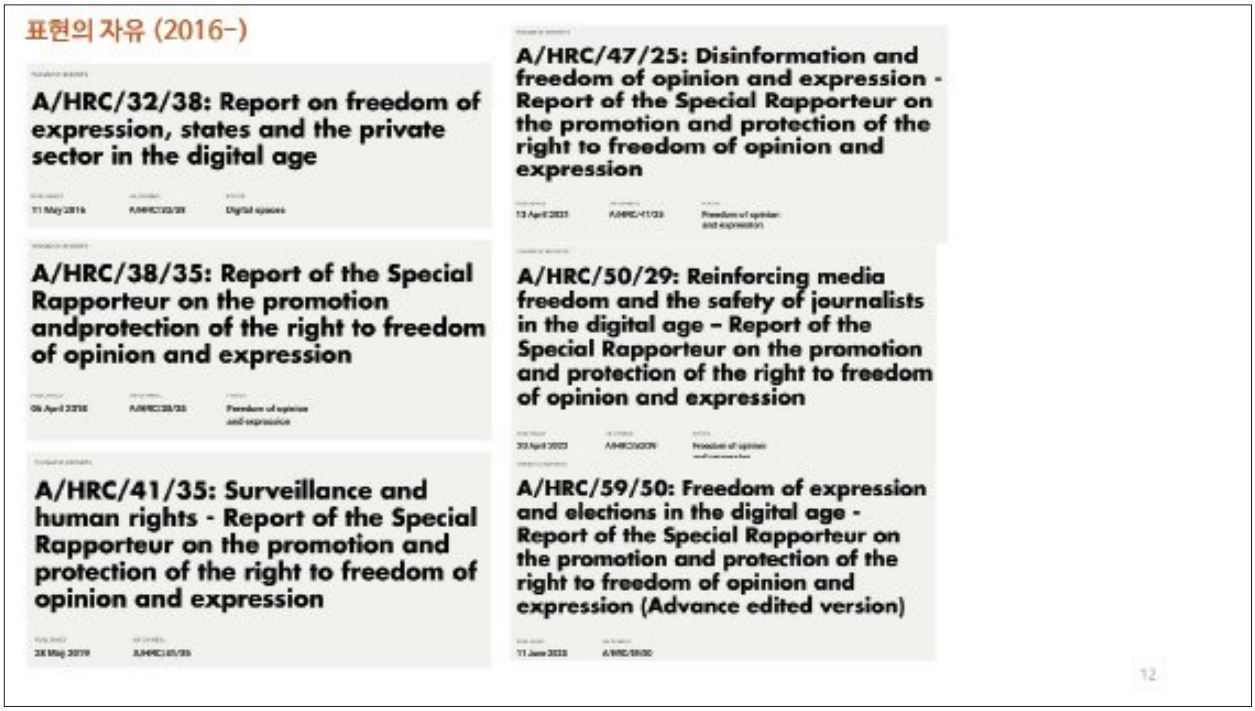
10

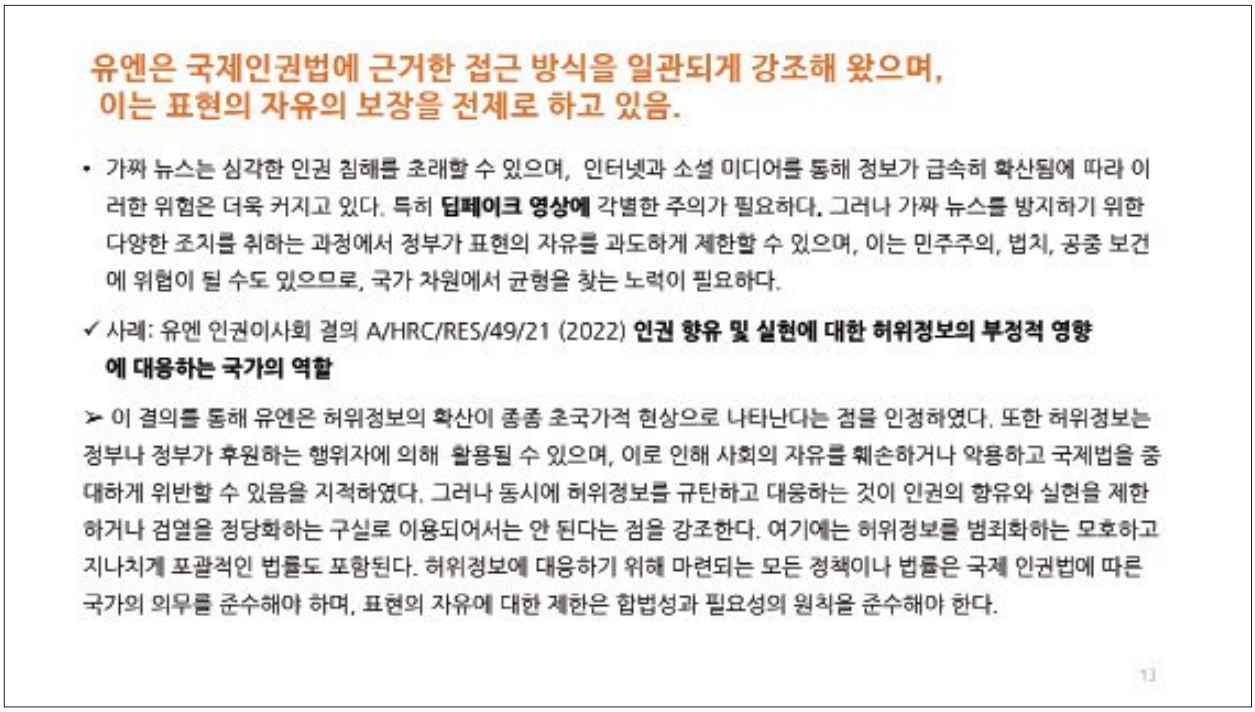
디지털 시대의 프라이버시권(2014-)

유엔 인권이사회는 2014년부터 '디지털 시대의 프라이버시권'을 다루어 왔으며, 초기에는 **국가 감시 및 인터넷 차단,** 사생활 보호와 표현의 자유 및 정보 접근에 관한 **협력 및 관할권 문제에** 대한 우려를 제기하였다. 또한 **디지털 격차로** 인한 교육, 식량, 의료 등 다양한 경제적·사회적·문화적 측면에 미치는 영향에 대한 우려를 표명하였다. 2020년 발표된 유엔 사무총장의 <디지털 협력을 위한 로드맵>에서는 **데이터. 사생활 보호. 디지털 신원. 역량 강화, 온라인 폭력** 관련 문제가 더욱 심각해졌음을 지적하며, 현재 및 미래의 국제 인권 규범이 온라인에서도 반드시 적용되어야 함을 강조했다.

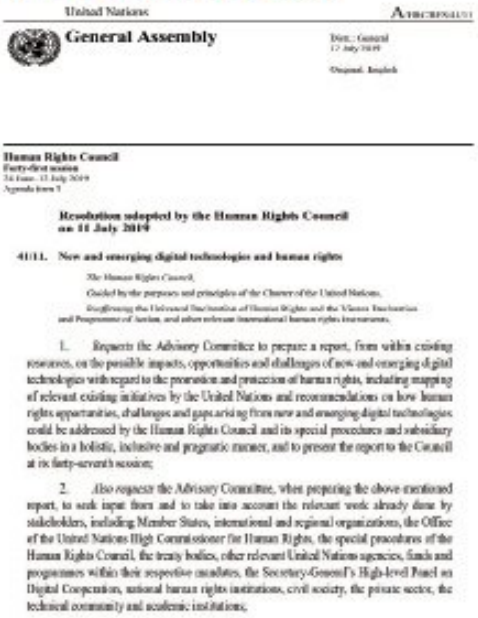
현재 유엔 인권이사회에서 디지털 기술과 인권과 관련해 가장 활발히 논의되는 문제는 **프라이버시권이다**. 이미 2014년에 감시와 모니터링과 관련된 인권 침해 문제를 제기하고, 이러한 문제를 식별,파악하기 위해 '디지털 시대의 사생활 보호권' 보고서가 발간되었다. 2018년부터는 디지털 및 신기술로 인해 발생하는 보다 구체적이고 다양한 형태의 인권 침해 사례가 검토되었으며, 2021년에는 인공지능 기술이 사생활 권리에 미치는 영향을 분석하였다. 유엔은 기존의 국제 인권 규범이 온라인 환경에도 적용될 수 있다고 보았으나, 온라인 환경에서만 적용 가능한 새로운 인권 규범 또는 **'디지털 권리'**의 필요성에 대해서는 여전히 논의가 계속되고 있다.

11


표현의 자유 (2016-)
A/HRC/32/38: Report on freedom of expression, states and the private sector in the digital age
A/HRC/38/35: Report of the Special Rapporteur on the promotion andprotection of the right to freedom of opinion and expression
A/HRC/41/35: Surveillance and human rights - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/47/25: Disinformation and freedom of opinion and expression - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/50/29: Reinforcing media freedom and the safety of journalists in the digital age – Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression
A/HRC/59/50: Freedom of expression and elections in the digital age - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Advance edited version)
12


유엔은 국제인권법에 근거한 접근 방식을 일관되게 강조해 왔으며,
이는 표현의 자유의 보장을 전제로 하고 있음.
• 가짜 뉴스는 심각한 인권 침해를 초래할 수 있으며, 인터넷과 소셜 미디어를 통해 정보가 급속히 확산됨에 따라 이러한 위험은 더욱 커지고 있다. 특히 딥페이크 영상에 각별한 주의가 필요하다. 그러나 가짜 뉴스를 방지하기 위한 다양한 조치를 취하는 과정에서 정부가 표현의 자유를 과도하게 제한할 수 있으며, 이는 민주주의, 법치, 공중 보건에 위협이 될 수도 있으므로, 국가 차원에서 균형을 찾는 노력이 필요하다.
✓ 사례: 유엔 인권이사회 결의 A/HRC/RES/49/21 (2022) 인권 향유 및 실현에 대한 허위정보의 부정적 영향에 대응하는 국가의 역할
➢ 이 결의를 통해 유엔은 허위정보의 확산이 종종 초국가적 현상으로 나타난다는 점을 인정하였다. 또한 허위정보는 정부나 정부가 후원하는 행위자에 의해 활용될 수 있으며, 이로 인해 사회의 자유를 훼손하거나 악용하고 국제법을 중대하게 위반할 수 있음을 지적하였다. 그러나 동시에 허위정보를 규탄하고 대응하는 것이 인권의 향유와 실현을 제한하거나 검열을 정당화하는 구실로 이용되어서는 안 된다는 점을 강조한다. 여기에는 허위정보를 범죄화하는 모호하고 지나치게 포괄적인 법률도 포함된다. 허위정보에 대응하기 위해 마련되는 모든 정책이나 법률은 국제 인권법에 따른 국가의 의무를 준수해야 하며, 표현의 자유에 대한 제한은 합법성과 필요성의 원칙을 준수해야 한다.
13

유엔 인권이사회 결의 (2019)



United Nations

General Assembly

Human Rights Council
Forty-first session

Resolution adopted by the Human Rights Council on 11 July 2019

41/11. New and emerging digital technologies and human rights

The Human Rights Council,

Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

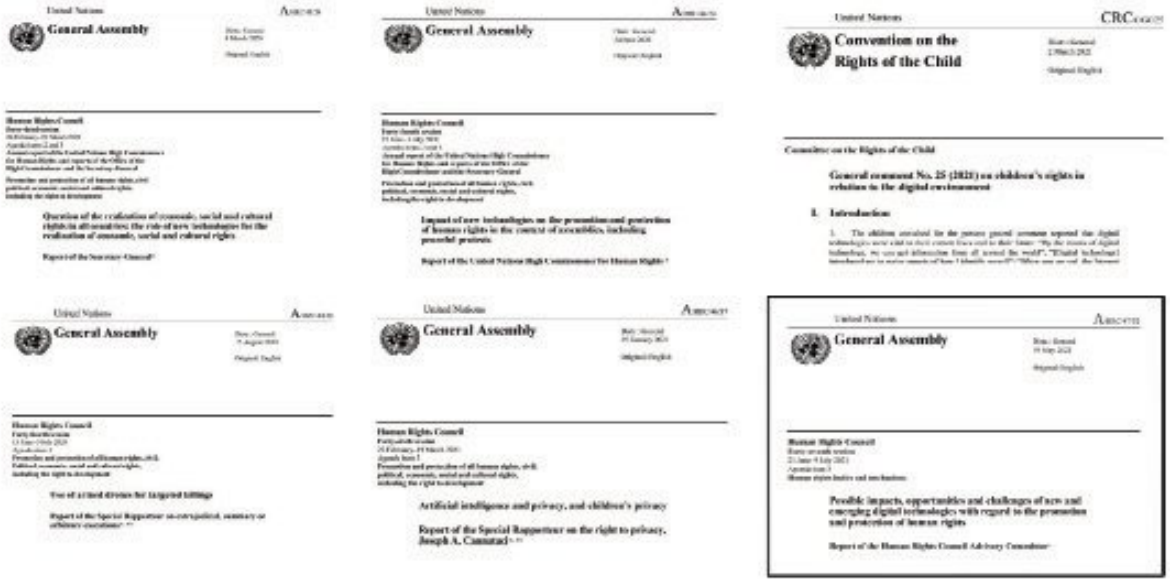
1. Requests the Advisory Committee to prepare a report, from within existing resources, on the possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights, including mapping of relevant existing initiatives by the United Nations and recommendations on how human rights opportunities, challenges and gaps arising from new and emerging digital technologies could be addressed by the Human Rights Council and its special procedures and subsidiary bodies in a holistic, inclusive and pragmatic manner, and to present the report to the Council at its forty-seventh session;

2. Also requests the Advisory Committee, when preparing the above-mentioned report, to seek input from and to take into account the relevant work already done by stakeholders, including Member States, international and regional organizations, the Office of the United Nations High Commissioner for Human Rights, the special procedures of the Human Rights Council, the treaty bodies, other relevant United Nations agencies, funds and programmes within their respective mandates, the Secretary-General's High-level Panel on Digital Cooperation, national human rights institutions, civil society, the private sector, the technical community and academic institutions;

1. 자문위원회가 기존 자원을 활용하여, 인권의 증진과 보호와 관련한 디지털 신기술의 잠재적 영향, 기회 및 도전과제에 관한 보고서를 준비할 것을 요청한다. 여기에는 유엔이 추진중인 관련 기존 이니셔티브를 검토하고, 인권과 관련한 기회, 도전, 격차를 신기술의 등장으로부터 도출하며, 인권이사회와 그 특별절차 및 보조기구들이 이를 총체적, 포용적, 실용적 방식으로 어떻게 다룰 수 있는지에 관한 권고사항을 포함해야 한다. 또한 이 보고서를 제47차 인권이사회에 제출할 것을 요청한다.

2. 또한 자문위원회가 상기 보고서를 준비함에 있어, 회원국, 국제 및 지역기구, 유엔 인권최고대표사무소, 인권이사회의 특별절차, 조약기구, 기타 관련 유엔 기구, 기금, 프로젝트(각 권한 범위 내에서), 사무총장의 디지털 협력 고위급 패널, 국가인권기구, 시민사회, 민간부문, 기술 커뮤니티 및 학계 등 이해관계자들이 이미 수행한 관련 작업을 반영하고 의견을 수렴할 것을 요청한다.

14



United Nations

General Assembly

Human Rights Council

Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights

Report of the Secretary-General

United Nations

General Assembly

Human Rights Council

Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests

Report of the United Nations High Commissioner for Human Rights

United Nations

Convention on the Rights of the Child

Committee on the Rights of the Child

General comment No. 25 (2021) on children's rights in relation to the digital environment

I. Introduction

United Nations

General Assembly

Human Rights Council

Use of armed drones for targeted killings

Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions

United Nations

General Assembly

Human Rights Council

Artificial intelligence and privacy, and children's privacy

Report of the Special Rapporteur on the right to privacy, Joseph A. Cannataci

United Nations

General Assembly

Human Rights Council

Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights

Report of the Human Rights Council Advisory Committee

15

유엔의 디지털 신기술과 인권에 관한 논의

- 유엔 내에서 이루어지고 있는 신기술과 인권에 관한 기존 논의는 다음과 같은 전제에 기반한다:

첫째, 기술이 중립적이기 때문에 기술 발전의 부정적 결과는 오로지 인간의 오용과 남용에서 비롯된 것이라는 주장은 현상을 지나치게 단순화하는 것이다. 신기술의 이용자 뿐만 아니라 신기술 그 자체도 인권 향유를 제한하고, 인권 정책에 영향을 미치며, 개인의 자유를 억압할 수 있다.

둘째, 신기술이 인권에 미치는 영향을 검토하기 위해서는 통합적이고 간학제적인 관점이 필요하다. 이는 모든 신기술이 처음부터 인권을 준수하도록 설계된 것은 아니기 때문이다. 신기술이 지닌 잠재적 인권 침해 요소에 충분한 주의를 기울여야 한다.

16

디지털 신기술에 의한 중대한 인권 침해

- 첫째, 신기술을 통한 개인정보의 과도한 데이터화는 필연적으로 사생활 침해 위험을 증대시킨다. 사생활 침해 문제는 다른 인권과도 밀접하게 연결되어 있으므로, 프라이버시권을 해치는 요소를 기술 발전의 불가피한 비용으로 용인해서는 안된다. 예컨대, 디지털 서비스의 데이터 처리 알고리즘은 매우 복잡하여, 일반 이용자가 이를 충분히 이해한 뒤 개인정보 활용에 대해 '정보에 입각한 동의'를 했다고 보기 어렵다.

- 둘째, 취약한 사이버 보안 시스템은 심각한 개인정보 침해로 이어질 수 있다. 사용자 데이터를 기반으로 한 공공 및 민간의 비즈니스·거버넌스 모델을 설계하고 운영함에 있어, 개인의 사생활 보장과 개인정보 유출 방지보다 다른 요인이 우선시되는 경우가 많다.

17

디지털 신기술에 의한 중대한 인권 침해

- 셋째, 디지털 시대에는 신기술을 통한 정보 확산 속도가 더욱 빨라진 반면, 정보 획득 비용은 상대적으로 낮아졌다. 그러나 역설적으로, 신뢰할 수 있고 출처가 명확한 정보와 잘못된 정보(misinformation)와 역정보(disinformation)를 구분하는 것은 더 어려워졌다. 인터넷은 미디어 콘텐츠의 생산 및 소비 방식에 엄청난 변화를 가져왔지만, 신기술의 발전으로 인해 정보의 신뢰성을 유지하고 진위를 판단하는 것이 오히려 더 어려워졌다.
- 넷째, 신기술의 발전은 혐오 표현의 급속한 확산을 가능하게 하였고, 이는 급진주의, 분리주의, 혐오 범죄 및 다양한 형태의 차별로 이어질 수 있다. 일부 디지털 미디어와 소셜 네트워크 서비스는 혐오 표현의 증가와 혐오적 사상의 확산에 기여해 왔다. 한편, 인공지능 기반 의사결정 시스템이 편향된 알고리즘을 바탕으로 설계될 경우(개발자의 의도와 무관하게) 차별적 결과가 발생할 수 있다.

18

디지털 신기술에 의한 중대한 인권 침해

- 다섯째, 인터넷이 주요한 의사소통 및 정보 접근 수단으로 자리 잡으면서, 정보 접근성이 부족한 취약 계층은 필연적으로 인권 침해 요소에 더 많이 노출된다. 문제는 향후 기술 발전이 이러한 비대칭적 정보 접근 양상을 더욱 심화시킬 가능성이 있다는 점이다. 이는 사회 내 기존 불평등을 악화시킬 뿐만 아니라 새로운 형태의 사회적 취약성과 취약 계층을 만들어 낼 수도 있다.
- 여섯째, 신기술은 모든 사람에 대한 무차별적 감시를 용이하게 하며, 이는 각국 정부의 불법적이고 자의적인 대규모 감시로 이어질 수 있다. 공공 질서나 공익을 위해 시행되는 감시 정책이라 하더라도, 적절한 인권 보호 장치가 마련되지 않으면 개인의 사생활을 부당하게 침해하는 행위로 쉽게 변질될 수 있다. 같은 맥락에서 디지털 공간은 언론의 자유, 정보 접근권, 표현의 자유, 집회 및 결사의 자유를 제한하는 데 이용될 위험을 내포하고 있다.
- 일곱째, 신기술은 딥페이크를 통한 성폭력, 성 착취, 온라인 괴롭힘, 금융정보 도용 등 전례 없는 새로운 범죄를 만들어냈다. 유엔 프라이버시권 특별보고관 역시 신기술이 성별 기반 폭력의 형태를 다양화하고 확대시켰다고 지적했다.

19

현행 국제인권 체제에서 디지털 신기술로 제기되는 다양한 인권 문제에 대한 대응의 이론적, 실천적 격차

- 첫째, 개념상의 한계가 존재한다. 지금까지 국제 인권 규범은 온라인이 아닌 오프라인 상황을 전제로 인권 침해 문제를 다루고 해결하기 위해 만들어져 왔기 때문에, 디지털 시대의 현실이 충분히 반영되지 못했다. 물론 이는 유엔이 신기술과 인권 문제를 해결하기 위한 최선의 방법으로 새로운 인권 조약이나 국제 협정을 채택하거나 기존의 문서를 개정해야 한다는 입장을 취한다는 뜻은 아니다.
- 전문성에 따른 격차도 존재한다. 신기술 분야 종사자들은 일반적으로 인권에 대한 이해가 부족하며, 반대로 인권 전문가들은 기술에 대한 이해가 부족할 수 밖에 없다. 이로 인해 이른바 인권 간 충돌(human rights trade-off)라는 역설적 상황이 발생할 수 있는데, 신기술이 특정 인권의 향유를 증진시키는 동시에 다른 인권을 침해하는 경우가 이에 해당한다. 인권 기반의 확립된 지침이나 규범이 없다면, 신기술 시스템 설계자와 기업 실무자들은 자신들에게 유리한 특정 인권만을 선택하고 보호할 위험이 있다. 그 결과, 개별 기업의 선호에 맞춘 제한적 범위의 인권만을 다루는 자율적 윤리 행동 강령이 사회에 확산 될 수 있으며, 실제로 그러한 움직임이 이미 나타나고 있다.

20

현행 국제인권 체제에서 디지털 신기술로 제기되는 다양한 인권 문제에 대한 대응의 이론적, 실천적 격차

- 둘째, 운영상의 격차를 들 수 있다. 신기술 발전과 이를 규제할 규범 사이에는 불가피한 시차가 존재한다. 새로운 규범을 만들기 위해서는 먼저 사회적 합의가 필요하다. 따라서 정부는 규범이 마련될 때까지는 민간 부문이 기존 인권 규범을 자발적으로 준수해 주기를 바랄 수밖에 없다. 이러한 운영상의 시차는 국제 거버넌스 수준에서도 다양한 문제를 야기한다. 신기술의 영향력과 파급력은 국제적이고 초국적이지만, 지금까지의 규제는 국가적 또는 지역적 차원에 국한되어 왔다. 실제로 국제기구와 개별 유엔 회원국이 신기술 관련 정책을 독립적으로 시행할 경우, 이들 정책사이에서 규제와 운영의 중복이 불가피하게 발생할 가능성이 크다. 따라서 이러한 중복을 방지하고 운영상의 격차를 좁히기 위해서는 개별 국가와 국제기구 간의 국제적 협의와 논의가 필수적이다.

21

현행 국제인권 체제에서 디지털 신기술로 제기되는 다양한 인권 문제에 대한 대응의 이론적, 실천적 격차

- 한편, 인권 보호에서 민간 부문의 역할이 확대됨에 따라 정부와의 관계에서 운영상의 격차가 발생할 수 있다. 인권 보호 및 증진에 대한 1차적 의무는 여전히 국가에 있지만, 2011년『유엔 기업과 인권 이행지침(UNGPs)』채택 이후 지난 10년간 민간 부문이 부담해야 할 인권 보호 의무와 역할에 대한 논의는 상당히 진정되어 왔다. 그러나 신기술을 기반으로 한 일부 혁신적인 비즈니스 모델은 법의 사각지대에 존재하거나 이를 악용하도록 설계되는 경우도 있어, 기술 기업을 포함한 민간 부문이 어떤 역할과 의무를 져야 하는지 신중한 검토가 필요하다.

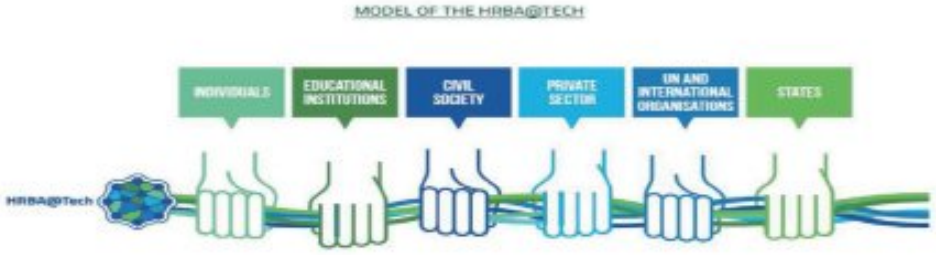
22

신기술에 대한 포괄적이고 통합적인 인권 기반 접근법

- 첫째, 기술에 대한 포괄적인 이해가 필요하다. 이는 개별 기술의 이해에 그치지 않고, 신기술과 데이터화 과정의 고유한 복잡성과 상호의존성을 종합적으로 파악하는 것을 의미한다. 다양한 유형의 기술 발전과 혁신 관련 인권 문제의 상호연결성을 충분히 이해해야 한다. 또한 기술 개발 과정 전반-설계, 구현, 실행부터 폐기 단계-에 이르기까지 발생할 수 있는 잠재적 인권 관련 문제들을 이해하는 것이 중요하다.
- 둘째, 인권에 대한 포괄적인 접근이 필요하다. 이는 기업과 기술 개발자들이 실질적으로 이해할 수 있는 공통 언어로 개별 인권 규범을 설명하고 표현해야 함을 의미한다. 첫 번째 요소가 인권 전문가들이 신기술을 이해하기 위해 노력해야 함을 뜻한다면, 두 번째 요소는 반대로 기술 개발자들이 인권의 언어와 내용을 이해해야 함을 의미한다. 이는 기술 개발 초기 단계부터 인권에 영향을 미칠 많은 핵심 결정이 개발자와 엔지니어에 의해 이루어지기 때문에 더욱 그러하다. 따라서 신기술 설계 및 개발 초기 단계부터 여성·장애인·아동 등 사회적 소수자의 권리를 포함한 시민적 권리와 사회적 권리를 보호하고 증진하기 위한 고민이 반드시 포함되어야 한다.

신기술에 대한 포괄적·통합적 인권 기반 접근법

- 셋째, 신기술과 인권 거버넌스 및 규제를 위해 개별 국가, 국제기구 및 모든 이해관계자의 포괄적 노력이 필요하다. 신기술이 인권에 미치는 영향은 다수 이해관계자 간 협력을 통해서만 효과적으로 규율될 수 있다. 따라서 신기술에 대한 인권 보호 체계를 효과적으로 실현하고 강화하며 동시에 지속 가능한 모니터링 체계를 구축하기 위해서는 민간 부문, 학계, 특히 시민사회와 같은 다양한 주체의 의미 있는 참여가 보장되어야 한다.



24

유엔 인권이사회 결의(2021)

United Nations A/HRC/RES/47/23

General Assembly

Human Rights Council

Resolution adopted by the Human Rights Council on 13 July 2021

47/23. New and emerging digital technologies and human rights

The Human Rights Council,

Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

3. Requests the Office of the High Commissioner to convene two expert consultations, to discuss the relationship between human rights and technical standard-setting processes for new and emerging digital technologies and the practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, and to submit a report thereon, reflecting the discussions held in an inclusive and non-prohibitive manner, to the Human Rights Council at its fiftieth and fifty-third sessions;

4. Also requests the Office of the High Commissioner, when preparing the above-mentioned expert consultations and reports, to seek input from and to take into account the relevant work already done by stakeholders from diverse geographic regions, including States, international and regional organizations, the Advisory Committee, the special procedures of the Human Rights Council, the treaty bodies, other relevant United Nations agencies, funds and programmes, including the International Telecommunication Union, other relevant standard development organizations, and the Office of the Envoy of Secretary-General on Technology, within their respective mandates, national human rights institutions, civil society, the private sector, the technical community and academic institutions;

3. 유엔 인권최고대표사무소는 두 차례 전문가 회의를 소집하여, 인권과 새로운 디지털 기술의 기술 표준 제정 과정 간의 관계, 그리고 기업과 인권 이행지침의 실제적 적용을 기술 기업 활동과 연계하여 논의하도록 요청한다. 또한 이러한 논의를 포괄적이고 포용적인 방식으로 반영한 보고서를 작성하여 제50차 및 제51차 인권이사회에 제출하도록 요청한다.

4. 또한 유엔 인권최고사무소가 위의 전문가 협의 및 보고서를 준비함에 있어, 다양한 지역에 속한 이해관계자들이 이미 수행한 관련 작업을 반영하도록 요청한다. 여기에는 국가, 국제 및 지역 기구, 자문위원회, 인권이사회의 특별절차, 조약기구, 국제전기통신연합을 포함한 기타 관련 유엔 기구·기금·프로그램, 기타 관련 표준 제정기구, 유엔 사무총장 기술특사실, 국가인권기구, 시민사회, 민간 부문, 기술 공동체 및 학계 등이 포함된다.

25

기업과 인권 논의와의 연계

- 유엔의 신기술과 인권에 관한 논의는 주로 연성 규범(soft law) 형태의 국제 규범 형성을 위한 기업과 인권 논의와 연계되어 진행되어 왔다.
- 특히 2019년 7월 30일, 유엔 인권이사회 기업과 인권 실무그룹은 디지털 기술에 관한 B-Tech 프로젝트를 시작하며, 디지털 기술 발전으로 비롯되는 인권 침해의 부정적 영향에 대응할 필요성을 제기하였다. 이 프로젝트의 주요 목적은 다양한 기업, 시민사회단체 및 정책 이해관계자의 참여를 통한 폭 넓은 협의와 연구를 통해 국가와 기업을 위한 실질적인 정보, 지침 및 권고안을 마련하는 것이다. B-Tech 프로젝트는 특히 다음 네 가지 전략적 중점 분야를 검토하고 있다:
  - 인권 위험에 대응할 수 있는 비즈니스 모델 마련
  - 인권 실사 및 최종 용도(End-Use)
  - 책임 및 구제
  - 스마트 믹스 방안

UNITED NATIONS HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER
B-Tech

26

기업과 인권 논의와의 연계

- 2022년 4월, "기업과 인권 이행지침의 기술 기업 활동에 대한 실제적 적용 " (A/HRC/50/56, 2022년 4월 21일)이 발표되어 보다 구체적인 이행 방안을 제시하였다.
- 보고서는 "기업과 인권 이행지침"에 명시된 국가의 보호 의무, 기업의 존중 책임, 구제 수단 접근성이 기술 산업에도 모두 적용되어야 함을 강조하고, 특히 이 분야의 기업들이 신기술 설계·개발·활용 과정에서 인권 실사(due diligence)를 수행하도록 적극 권장한다. 또한 디지털 격차와 성평등을 고려하여 취약 계층에 대한 관심을 더욱 기울일 필요성을 강조하였다.

27

유엔 인권이사회 결의안(2023)

United Nations

General Assembly

Human Rights Council

Resolution adopted by the Human Rights Council on 14 July 2023

51/21. New and emerging digital technologies and human rights

5. Requests the Office of the High Commissioner to prepare a report, in consultation with States, mapping the work and recommendations of the Human Rights Council, the Office of the High Commissioner, the treaty bodies and the special procedures of the Human Rights Council in the field of human rights and new and emerging digital technologies, including artificial intelligence, as well as identifying gaps and challenges and making recommendations on how to address them, while giving due consideration to the United Nations system-wide work on new and emerging digital technologies, and to present the report to the Council at its fifty-sixth session, to be followed by an interactive dialogue;

6. Requests the United Nations High Commissioner for Human Rights to expand the capacities within the Office of the High Commissioner, to advance human rights in the context of new and emerging digital technologies, including at the regional level, and to provide advice and technical assistance to States, upon their request, on issues concerning human rights and new and emerging digital technologies, including artificial intelligence, and to, as appropriate, all relevant United Nations organizations and bodies;

7. Requests the Office of the High Commissioner to continue to work on the practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, including by convening an expert consultation, including with States and business enterprises, including technology companies, civil society and academia, to discuss challenges, good practices and lessons learned in applying the Guiding Principles to the activities of technology companies, including activities relating to artificial intelligence, and to submit a report thereon to the Human Rights Council at its fifty-ninth session;

5. 인권 최고대표사무소가 국가들과 협의하여, 인권이사회, 인권최고대표사무소, 조약기구, 인권이사회의 특별절차가 인공지능을 포함한 새로운 디지털 기술 분야에서 수행한 작업과 권고를 정리, 분석하고 신기술 관련 격차와 도전 과제를 식별하며, 이를 어떻게 해결 할 수 있을지에 대한 권고를 마련하도록 요청한다. 또한 유엔 시스템 전체 차원에서의 신기술 관련 활동을 충분히 고려하여 보고서를 작성하여 제56차 인권이사회에 제출하며, 이후 상호 대화로 이어지도록 한다.

6. 인권최고대표가 인공지능을 포함한 새로운 기술의 맥락에서 인권 증진을 위해, 지역적 차원을 포함하여 인권최고대표사무소의 역량을 강화하고, 회원국의 요청이 있는 경우, 인권 및 새로운 디지털 기술과 관련된 문제에 대해 자문과 기술적 지원을 제공하도록 요청하며, 필요에 따라 모든 관련 유엔 기구 및 기관들과 협력하도록 한다.

7. 인권최고대표사무소가 '기업과 인권 이행지침'을 기술 기업 활동에 실제로 적용하는 작업을 지속하도록 요청한다. 이를 위해 전문가 협의를 소집하고, 국가와 기술 기업을 포함한 기업, 시민사회, 학계가 참여하여 인공지능을 포함한 기술 기업 활동에서 직면하는 도전과제, 모범사례, 교훈을 논의하며, 그 결과를 보고서로 작성해 제59차 인권이사회에 제출하도록 한다.

28

인공지능과 인권 관련 논의

- 지금까지 국제사회 논의의 특성을 살펴보면, 인공지능 시스템의 운영 및 규제와 관련된 법적 쟁점들은 거의 전적으로 기존 국제 인권법 체계를 기반으로 논의되고 있다. 그러나 기존 법체계가 인공지능 윤리와 둘러싼 모든 문제를 충분히 포괄할 수 있는지에 대해서는 명확한 답이 없다. 인용되는 국제 인권법의 범위는 국제조약 뿐 아니라 선언, 결의, 권고, 지침까지 확장되며, 해당 국제 문서의 법적 효력과는 무관하다. 또한 기술적·재정적 역량과 사회적 환경은 국가 별 차이가 크기 때문에 다양한 권고사항을 국내적으로 이행하는 과정에서 격차가 불가피하게 발생한다. 구체적인 이행 과정에서는 국가별로 선택적 수용이나 수정 현상이 발생할 수 있으며, 이 경우 인공지능 윤리 관련 권고사항과 지침은 본래 의도된 목표를 달성하기보다 불평등하고 단편적이며 희석된 결과로 이어질 수 있다. 이는 구속력이 없는 권고사항과 지침의 자발적 이행 과정이 지니는 한계를 보여준다.

29

인공지능과 인권 관련 논의

- 인공지능이 일상생활에 깊숙이 자리 잡은 상황에서 인권 침해를 예방하고, 침해가 발생했을 때 대응할 수 있는 체계를 구축해야 한다는 데는 이견이 없다. 그러나 사용자의 알 권리와 인공지능 개발자의 영업비밀 보호 권리가 충돌할 때 이를 조정할 수 있는 제도적 장치가 필요하다. 예를 들어, 제약업계는 전통적으로 신약 출시 전 위험성을 사전에 진단하고 일반 대중 사용에 안전하다고 판단될 때만 시장에 출시할 수 있도록 하는 허가 제도를 채택해 왔다. 마찬가지로 AI 소프트웨어의 경우, 이미 개발되어 출시된 이후 점검하는 방식이 아니라 개발 단계에서 시험을 거쳐 문제 발견 시 즉시 수정할 수 있도록 하는 'AI 소프트웨어 사전 인증 제도'를 고려할 수 있다. 또한 신기술이 초래할 수 있는 부정적 영향이 광범위하므로, 잠재적 영향을 예방하기 위해 가능한 이른 시점에 인권 영향 평가를 실시해야 한다. 더 나아가 신기술을 기획·개발하는 기업뿐 아니라 특정 신기술을 도입·활용하는 타 산업 분야 기업도 인권 영향 평가를 의무적으로 시행해야 한다.

30

보도자료

인공지능 개발과 활용에 있어서 인권 보호를 위해 인권영향평가 실시 필요

- 인권위 「인공지능 인권영향평가 도구」 마련, 「인공지능 인권 가이드라인」(22)에 이은 구체적인 인권 보호 기준 제시 -

□ 국가인권위원회(위원장 송두환, 이하 '인권위')는 2024년 5월 23일 과학기술정보통신부장관에게, 인공지능의 개발과 활용으로 인한 인권침해 및 차별이 발생하지 않도록 관련 정책 · 사업계획의 수립 · 이행 시에 「인공지능 인권영향평가 도구」를 활용하고, 인권영향평가가 법제화하기 전까지 공공기관이 직접 개발하거나 활용하는 모든 인공지능과 민간 부문의 고위험 인공지능에 대한 사용자 인권영향평가가 수행될 수 있도록 이 평가 도구를 보급할 필요가 있다는 의견을 표명하였다.

□ 인공지능 기술의 발전과 확산이 우리 사회에 긍정적 변화를 가져올 것이라는 기대가 높아지는 반면, 인공지능으로 인한 인권침해와 차별 등 부정적 영향에 대한 우려도 점점 커지고 있다.

○ 인공지능 시스템을 학습시 불투명성의 악순환과 같으로 인하여 사후적인 피해구제나 책임기 어렵기 때문에 인공지능의 개발 및 활용에 있어 사전예방적 인권영향평가의 도입 및 실시가 필요하다.

평가 도구 구성

구분	단계	상세목차 및 주요 내용	문항수	면수
평가 개요	-	가. 인공지능 인권영향평가 도구의 성격 및 개요		1
		나. 인공지능 인권영향평가의 대상		1
		다. 인공지능 인권영향평가 시기		2
		라. 인공지능 인권영향평가 수행 주체		3
		마. 인공지능 인권영향평가의 절차		4
		바. 인공지능 인권영향평가의 수행 과정		7
		사. 인공지능 인권영향평가 도구만 활용 시나리오		8
평가 도구	총 4단계 73개 문항			
	1단계	[계획 및 준비]	10	11
		가. 인권영향평가 계획	10	11
		나. 사전 조사	0	10
	2단계	[분석 및 평가]	56	10
		가. 인공지능 기술과 권리를 영향 분석 및 평가	27	10
		(1) 개인정보보호	1	10
		(2) 데이터 편향	6	10
		(3) 알고리즘의 설명과 투명성	3	17
		(4) 차별금지	2	18
		(5) 설명가능성과 투명성	5	19
		(6) 자동적 절차와 인간의 개입	4	20
		(7) 보안	3	21
		(8) 접근성	2	21
		(9) 리터러시	1	22
		나. 인권에 미치는 영향 및 심각성	7	22
		(1) 영향을 입는 인권	3	22
		(2) 인권에 미치는 영향의 심각성	4	23
	3단계	[개선 및 구제]	11	25
		가. 원칙	1	25
		나. 절차	4	25
		다. 구제	4	26
		라. 이해관계자와의 의견수렴 및 협의	2	27
	4단계	[평가 및 환류]	6	29
		가. 인공지능 시스템의 주요 요소의 공개	1	29
		나. 인권영향평가 결과 공개	2	29
		다. 사후 모니터링	2	30
		라. 인권영향평가에 대한 점검	1	30
		마. 인권영향평가의 재수행	0	31
부록	인공지능 인권영향평가 지표 구성 목적과 개발자에 대한 질의 목록			32

유엔 인권이사회 결의안 (2025)

United Nations

General Assembly

Human Rights Council

Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

[...] draft resolution

[...]. New and emerging digital technologies and human rights

The Human Rights Council,

6. *Requests* the Office of the United Nations High Commissioner for Human Rights to expand its work on United Nations system-wide promotion, coordination and coherence on matters related to human rights in new and emerging digital technologies and, as part of this, to convene regular meetings, in a virtual format, of United Nations human rights mechanisms and relevant United Nations entities working on digital technology issues, to exchange information, improve coordination and reduce duplication;

7. *Also requests* the Office of the High Commissioner to prepare an analytical study, building on its previous report mapping the existing work of the Human Rights Council and the treaty bodies, outlining and clarifying States' obligations under international human rights law, as well as relevant norms and commitments, and the human rights responsibilities of business enterprises in line with the Guiding Principles on Business and Human Rights, across the life cycle of new and emerging digital technologies, identifying developments, gaps and recommendations on application and implementation, and to present the report to the Council at its sixty-second session;

6. 인권최고대표사무소에게 유엔 시스템 전반에서 새로운 디지털 기술과 관련된 인권 사안에 관한 촉진·조정·조화를 강화하도록 요청하며, 이를 위해 유엔 인권 메커니즘과 디지털 기술 문제를 다루는 관련 유엔 기구들이 정기적으로 화상회의를 개최하여 정보를 교환하고, 협조 체계를 강화하며, 중복을 줄이도록 한다.

7. 또한 인권최고대표사무소에게 이전 보고서를 토대로 분석적 연구를 준비하도록 요청한다. 여기에는 인권이사회와 조약기구의 기존 작업을 정리·분석하고, 국제인권법상 국가의 의무를 개괄하고 명확히 하며, 관련 규범과 약속, 그리고 '기업과 인권 이행지침'에 따라 기업이 져야 할 인권 책임을 포함한다. 새로운 디지털 기술의 생애주기 전반에 걸친 발전 과정에서 발생하는 문제와 격차를 파악하고, 적용 및 이행에 관한 권고를 마련하여 제62차 인권이사회에 보고하도록 한다.

32

유엔 인권이사회 결의안 (2025)

United Nations

General Assembly

Human Rights Council

Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

[...] draft resolution

[...]. New and emerging digital technologies and human rights

8. *Further requests* the Office of the High Commissioner to convene a multi-stakeholder intersessional meeting, ahead of the sixty-fourth session of the Human Rights Council, utilizing the margins of other scheduled meetings, inviting the participation of States, as well as United Nations mechanisms, bodies and specialized agencies, funds and programmes, intergovernmental organizations, and mechanisms working in the field of human rights and new and emerging technologies, national human rights institutions and other relevant bodies, digital technology business enterprises, the technical community, academics and experts, as well as non-governmental organizations in the field of new and emerging digital technologies, in order:

(a) To provide a space for sharing experiences, challenges, good practices and lessons learned in realizing a holistic, inclusive and comprehensive approach to the development and implementation of national legislation and policies relevant to digital technologies, and in respecting and promoting human rights and principles of international human rights law throughout the technology life cycle;

(b) To consider the above-mentioned analytical study and discuss further steps to improve the implementation of the obligations and commitments of States under international human rights law, and the responsibilities of business enterprises throughout the life cycle of new and emerging digital technologies, including through, upon request and within existing resources, the Human Rights Advisory Service on Digital Technologies;

(c) To promote United Nations human rights system outputs relating to new and emerging digital technologies in order to improve the implementation of relevant recommendations at the national level;

(d) To submit a summary report thereon to the Human Rights Council at its sixty-fourth session;

8. 인권최고대표사무소가 제64차 인권이사회 회기 전에 다자 이해관계자 회의를 개최하도록 추가로 요청한다. 이 회의는 다른 예정된 회의와 연계하여 열리며, 회원국 뿐 아니라 유엔의 인권 메커니즘, 국가인권기구, 관련 기관, 디지털 기술 기업, 기술 공동체, 학계 및 전문가, 그리고 새로운 디지털 기술 분야의 비정부기구들을 초청하여 진행된다. 이 회의의 목적은 다음과 같다.

a) 디지털 기술 관련 국내 법률과 정책의 개발 이행과정에서, 그리고 기술 생애주기 전반에 걸쳐 국제인권법 원칙을 존중하고 인권을 증진하는 총체적·포용적·종합적 접근을 실현 하는데 있어, 경험·과제·모범 사례 교훈을 공유할 수 있는 장을 제공하는 것.

b) 앞서 언급된 분석 연구를 검토하고, 국제인권법상 국가의 의무와 약속, 그리고 새로운 디지털 기술의 생애주기 전반에 걸쳐 기업이 져야 할 인권 책임의 이행을 개선하기 위한 추가적 조치를 논의하는 것. 요청이 있는 경우 기존 자원을 활용해 '디지털 기술 인권 자문 서비스'를 제공할 수 있다.

c) 새로운 디지털 기술과 관련된 유엔 인권 시스템 산출물을 촉진하여 국가 차원에서 권고의 이행을 개선하는 것.

d) 요약보고서를 작성하여 제64차 인권이사회에 제출하는 것

33

디지털 신기술과 인권에 관한 유엔 논의

- 유엔 인권이사회는 자문위원회에 다양한 분야에서 나타나는 새로운 기술의 인권적 함의를 검토하는 보고서를 준비할 것을 요청하였다.

현재 임무

→ 군사 분야의 새로운 디지털 기술이 인권에 미치는 영향
→ 허위정보가 인권의 향유와 실현에 미치는 영향
→ 기술 매개 성폭력
→ 인권의 완전한 향유에 대한 플라스틱 오염의 영향
→ 인공지능 시스템이 거버넌스에 미치는 영향
→ 일방적 강제조치가 건강권에 미치는 부정적 영향
→ 경제적 사회적 문화적 권리의 국내 법 집행을 통한 사회정의
→ 뉴로테크놀로지에 관한 인권 지침
→ 부패가 인권의 향유에 미치는 부정적 영향

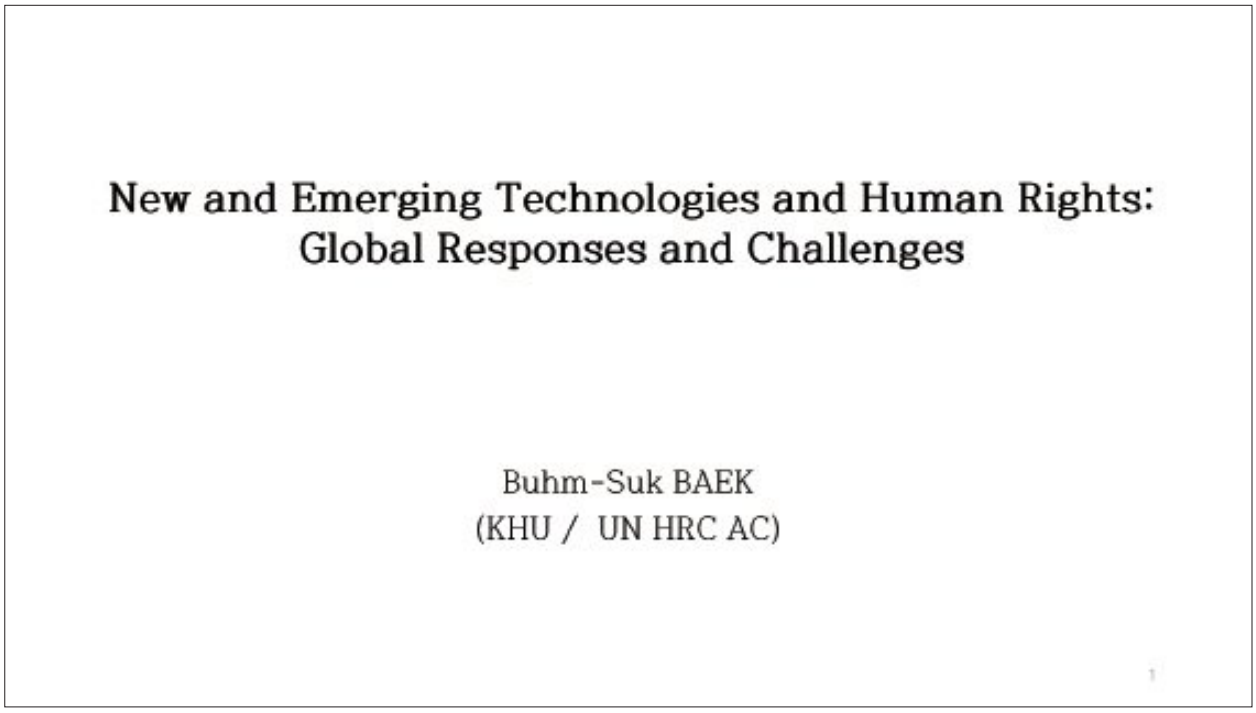
과거 임무 및 성과

→ 뉴로테크놀로지와 인권
→ 인권 향유를 위한 기후 보호 신기술의 영향
→ 인종 정의와 평등 증진
→ 디지털 신기술과 인권

34

New and Emerging Technologies and Human Rights: Global Responses and Challenges

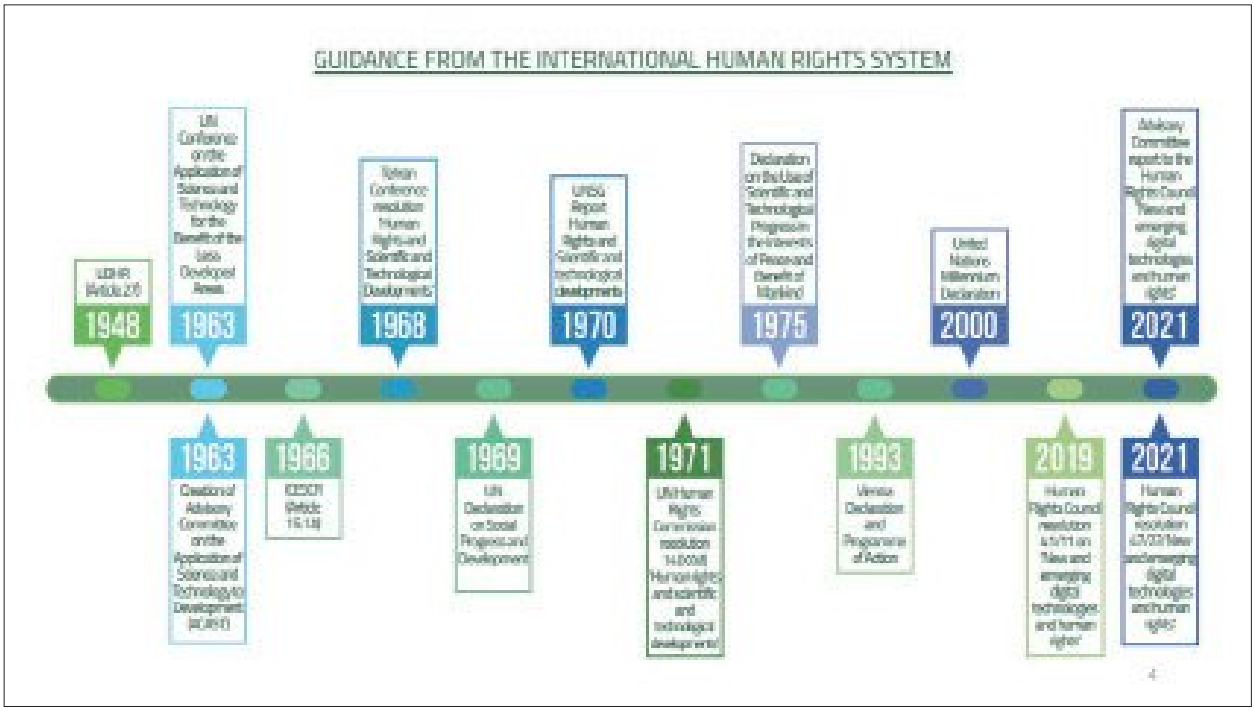
BAEK Buhm-suk | Professor, School of Law, Kyunghee University, Member, UN Human Rights Council Advisory Committee



- According to the UN Institute for Disarmament Research (UNIDIR) Cyber Policy Portal (https://cyberpolicyportal.org/), as of Aug. 2025, among the 30+ national position papers published to date, including those from the African Union (AU, with 55 African countries participating), Estonia, Canada, and Republic of Korea, 25 countries have addressed the application of international human rights law in cyberspace and emphasized the need for harmonization between digital technology and human rights.
- *"Respect for human rights constitutes a well-established obligation under international human rights law, and this obligation applies equally in the context of cyberspace. The Republic of Korea affirms that international human rights law applies online in the same manner as it does offline. Individuals are entitled to enjoy the same human rights in relation to cyber activities as they do in any other domain. Fundamental rights — including the right to privacy, freedom of expression, access to information, protection against discrimination, and the prevention of hate speech—must be protected and upheld in cyberspace for all individuals, including women and socially vulnerable groups."* **(National Position of the Republic of Korea on the Application of International Law in Cyberspace 2025)**

- Global Digital Compact (2024)

8(C). This Compact is anchored in international law, including international human rights law. All human rights, including civil, political, economic, social and cultural rights, and fundamental freedoms, must be respected, protected and promoted online and offline. Our cooperation will harness digital technologies to advance all human rights, including the rights of the child, the rights of persons with disabilities and the right to development.



- Universal Declaration of Human Rights Article 27

Article 27

Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.

Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.

- International Covenant on Economic, Social and Cultural Rights (ICESCR)

Article 15

The States Parties to the present Covenant recognize the right of everyone:(b) To enjoy the benefits of scientific progress and its applications.

Debate on the UN Human Rights Commission Resolution on Human Rights and Scientific and Technological Developments (1971)

The debate surrounding the relationship between human rights and technology has existed for a long time since the emergence of modern international human rights law. The discussion of the dual use (opportunity and challenges) aspects of new technologies is by no means a new discourse. Digital technology itself is not new, but over the years, exponential increases in performance have enabled the execution of complex tasks at faster speeds and scales than ever before. This vast availability of data and the growth of internet and broadband networks have led to the emergence and utilization of digital new technologies based on the internet and broadband networks, sometimes called the Fourth Industrial Revolution.

UN Discussions on Digital New Technologies and Human Rights

- Over the past 20 years, under the UN-centered international human rights mechanism system, various resolutions focusing on digital new technologies have been adopted, special procedures reports have been submitted, and general comments and recommendations from treaty bodies have been published, particularly focusing on the impact on specific rights such as the right to privacy, freedom of assembly and association, freedom of expression, children's rights, and social rights.
- In particular, the UN Human Rights Council first adopted a resolution on new technologies and human rights (A/HRC/RES/41/11) in 2019, and according to this resolution, the Advisory Committee submitted a report in 2021 on "Possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights."

UN Discussions on Digital New Technologies and Human Rights

UN Human Rights Council Resolutions
Promotion and Protection of All Human Rights, Civil, Political, Economic, Social and Cultural Rights, including the Right to Development: freedom of opinion and expression A/HRC/RES/17/16 (December 17, 2009)
The promotion, protection and enjoyment of human rights on the Internet A/HRC/RES/20/8 (July 18, 2012) The promotion, protection and enjoyment of human rights on the Internet A/HRC/RES/26/13 (July 14, 2014) The right to privacy in the digital age A/HRC/RES/28/16 (April 1, 2015)
Rights of the child: information and communication technologies and child sexual exploitation A/HRC/RES/31/7 (April 30, 2016) The promotion, protection and enjoyment of human rights on the Internet A/HRC/RES/32/13 (July 18, 2016) The right to privacy in the digital age A/HRC/RES/34/7 (April 7, 2017)
The right to privacy in the digital age A/HRC/RES/37/2 (April 6, 2018) The promotion, protection and enjoyment of human rights on the Internet A/HRC/RES/38/7 (July 17, 2018)
New and emerging digital technologies and human rights A/HRC/RES/41/11 (July 13, 2019) The right to privacy in the digital age A/HRC/RES/42/15 (October 7, 2019)

New and emerging digital technologies and human rights A/HRC/RES/47/23 (July 16, 2021)
The promotion, protection and enjoyment of human rights on the Internet A/HRC/RES/47/16 (July 26, 2021)
Right to privacy in the digital age A/HRC/RES/48/4 (October 13, 2021) Role of States in countering the negative impact of disinformation on the enjoyment and realization of human rights A/HRC/RES/49/21 (April 8, 2022) New and emerging digital technologies and human rights A/HRC/RES/52/29 (July 16, 2023)
Right to privacy in the digital age A/HRC/RES/54/21 (October 16, 2023) Safety of the child in the digital environment A/HRC/RES/56/8 (July 11, 2024) Promotion, protection and enjoyment of human rights on the Internet A/HRC/RES/57/29 (October 14, 2024)

UN Discussions on Digital New Technologies and Human Rights

Report of the OHCHR
The right to privacy in the digital age A/HRC/27/37 (30 June 2014) Information and communications technology and child sexual exploitation A/HRC/31/34 (1 December 2015) Promotion, protection and enjoyment of human rights on the Internet: ways to bridge the gender digital divide from a human rights perspective A/HRC/35/9 (5 May 2017) The right to privacy in the digital age A/HRC/36/29 (3 August 2018) Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights A/HRC/43/29 (4 March 2020) Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests A/HRC/44/24 (24 June 2020) The right to privacy in the digital age A/HRC/48/31 (13 September 2021) Systems and data collection under article 31 of the Convention on the Rights of Persons with Disabilities A/HRC/51/17 (3 December 2021) The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of Technology Companies A/HRC/50/56 (25 April 2022) Internet shutdowns: trends, causes, legal implications and impacts on a range of human rights – Report of the Office of the United Nations High Commissioner for Human Rights A/HRC/50/55 (13 May 2022) The right to privacy in the digital age A/HRC/51/17 (4 August 2022) Human rights and technical standard-setting processes for new and emerging digital technologies A/HRC/53/42 (9 June 2023) Panel discussion on the most efficient ways of upholding good governance to address the human rights impacts of the various digital divides A/HRC/55/50 (15 December 2023)

UN General Assembly Resolution
The right to privacy in the digital age A/RES/68/167 (21 January 2014) The right to privacy in the digital age A/RES/69/166 (18 February 2015) The right to privacy in the digital age A/RES/71/199 (25 January 2017) The right to privacy in the digital age A/RES/73/179 (21 January 2019) The right to privacy in the digital age A/RES/75/176 (28 December 2020) The right to privacy in the digital age A/RES/77/211 (15 January 2023)

The Right to Privacy in the Digital Age (2014-)

The UN Human Rights Council has been addressing 'privacy rights in the digital age' since 2014, initially raising concerns about **state surveillance and internet shutdowns, cooperation and jurisdiction issues** regarding privacy, freedom of expression and access to information, and expressing concerns about the impact on various economic, social and cultural aspects, including education, food, healthcare due to **the digital divide**. The UN Secretary-General's Roadmap for Digital Cooperation, published in 2020, highlighted that the seriousness of issues related to **data protection and privacy, digital identity, capacity building and online violence** has become more prominent, and emphasized that current and future international human rights norms must be applied online as well.

Currently, the most actively discussed human rights issue in the UN Human Rights Council regarding digital technology and human rights is **privacy rights.** As early as 2014, 'The Right to Privacy in the Digital Age' report was published to raise and detect human rights violation issues related to surveillance and monitoring. Since 2018, more specific and diverse forms of human rights violations caused by digital and new technologies have been examined, and in 2021, the impact of artificial intelligence technology on privacy rights was analyzed. The UN has also determined that existing international human rights norms can be applied online as well, but whether new human rights norms or **'digital rights'** that can only be applied online exist continues to be discussed.

The Freedom of Expression (2016-)

A/HRC/32/38: Report on freedom of expression, states and the private sector in the digital age

A/HRC/38/35: Report of the Special Rapporteur on the promotion andprotection of the right to freedom of opinion and expression

A/HRC/41/35: Surveillance and human rights - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

A/HRC/47/25: Disinformation and freedom of opinion and expression - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

A/HRC/50/29: Reinforcing media freedom and the safety of journalists in the digital age – Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

A/HRC/59/50: Freedom of expression and elections in the digital age - Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Advance edited version)

12

The UN has consistently emphasized an approach to disinformation based on international human rights law, premised on ensuring freedom of expression.

- Fake news can cause serious human rights violations. This risk is increasing as information can spread rapidly through the internet and social media. Special attention should be paid to **deep fake videos.** However, while taking various measures to prevent fake news, governments may excessively restrict freedom of expression, which could also threaten democracy, rule of law, and public health, so efforts to find a balance at the national level are necessary.

✓ For example: UN Human Rights Council Resolution A/HRC/RES/49/21 (2022) **Role of States in countering the negative impact of disinformation on the enjoyment and realization of human rights**

➢ Through this resolution, the UN recognized that the spread of disinformation is often a transnational phenomenon that can be used by governments or government-sponsored actors, which can undermine or exploit the freedoms of society and accompany serious violations of international law. However, it also emphasized that condemning and responding to disinformation should not be used as a pretext to restrict the enjoyment and realization of human rights or to justify censorship. This includes vague and overly broad laws that criminalize disinformation. All policies or laws undertaken to respond to disinformation must comply with state obligations under international human rights law, and restrictions on freedom of expression must comply with the principles of legality and necessity.

13

UN Human Rights Council resolutions (2019)

United Nations

General Assembly

A/HRC/41/L.11

Distr.: General
17 July 2019

Original: English

Human Rights Council
Forty-first session
24 June–12 July 2019
Agenda item 3

Resolution adopted by the Human Rights Council on 11 July 2019

41/11. New and emerging digital technologies and human rights

The Human Rights Council,

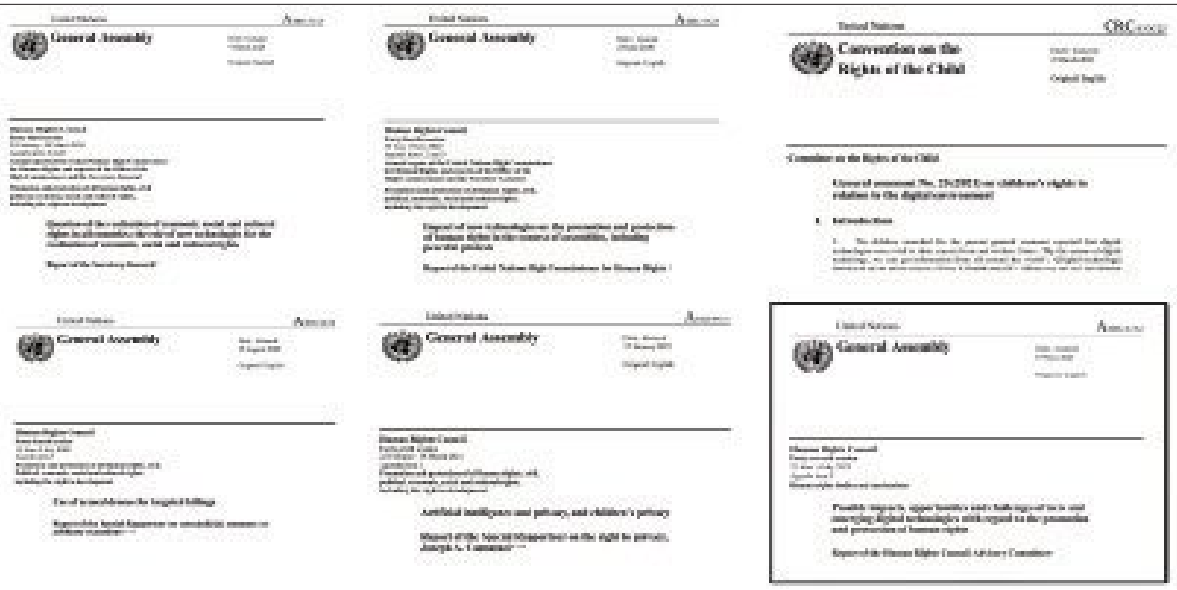
Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

1. *Requests* the Advisory Committee to prepare a report, from within existing resources, on the possible impacts, opportunities and challenges of new and emerging digital technologies with regard to the promotion and protection of human rights, including mapping of relevant existing initiatives by the United Nations and recommendations on how human rights opportunities, challenges and gaps arising from new and emerging digital technologies could be addressed by the Human Rights Council and its special procedures and subsidiary bodies in a holistic, inclusive and pragmatic manner, and to present the report to the Council at its forty-seventh session;

2. *Also requests* the Advisory Committee, when preparing the above-mentioned report, to seek input from and to take into account the relevant work already done by stakeholders, including Member States, international and regional organizations, the Office of the United Nations High Commissioner for Human Rights, the special procedures of the Human Rights Council, the treaty bodies, other relevant United Nations agencies, funds and programmes within their respective mandates, the Secretary-General's High-level Panel on Digital Cooperation, national human rights institutions, civil society, the private sector, the technical community and academic institutions;

14



15

UN Discussions on Digital New Technologies and Human Rights

- The existing discussions on new technologies and human rights within the UN are premised on the following:

First, the argument that negative consequences of technological development are purely caused by human misuse and abuse of technology because technology is neutral oversimplifies the phenomenon. Not only users of new technologies, but also new technologies themselves can limit the enjoyment of human rights, influence human rights policies, and suppress individual freedom.

Second, an integrated and interdisciplinary perspective is needed to examine the impact of new technologies on human rights. This is because not all new technologies are designed to be compliant with human rights from the outset. Sufficient attention needs to be paid to the potential human rights violation elements that new technologies possess.

Major Human Rights Violation by Digital New Technologies

- First, excessive datafication of personal information through new technologies inevitably increases the risk of privacy violations. Privacy violation issues are closely connected with other human rights, so elements that harm privacy rights should not be easily tolerated as inevitable costs of technological development. For example, data processing algorithms through digital services are very complex, making it difficult to consider that ordinary users fully understand them and have given informed consent to the use of personal information.
- Second, poor cybersecurity systems can lead to serious privacy violations. In designing and operating public and private business and governance models based on user data, the main concern is not in ensuring individual privacy rights and preventing personal information exposure.

Major Human Rights Violation by Digital New Technologies

- Third, the speed of information dissemination has become faster in the digital age through new technologies, while the cost of information acquisition has become relatively lower. However, paradoxically, distinguishing misinformation and disinformation from credible, clearly sourced information has become more difficult. The internet has brought tremendous changes to how media content is produced and experienced, but maintaining information reliability and determining authenticity has become relatively more difficult due to new technological developments.
- Fourth, developments in new technologies enable rapid spread of hate speech, which can cause radicalism, separatism, hate crimes and various forms of discrimination. Some digital media and social network services have contributed to increased hate speech and dissemination of hateful ideas. Meanwhile, if AI-powered decision-making systems are designed based on biased algorithms (regardless of whether this was intended by developers), discriminatory results can occur.

Major Human Rights Violation by Digital New Technologies

- Fifth, as the internet becomes a major means of communication and information access, vulnerable groups lacking information accessibility are inevitably more exposed to potential human rights violation elements. The problem is that future technological developments are likely to intensify such asymmetric information enjoyment patterns. This can worsen existing inequalities in society and even create new forms of social vulnerability and marginalized groups.
- Sixth, new technologies facilitate indiscriminate monitoring of entire populations, which can lead to illegal and arbitrary mass surveillance by individual governments. Even surveillance policies implemented for public order or public welfare can easily become acts that unduly violate individual privacy if appropriate human rights safeguards are not in place. In the same context, digital spaces pose risks of being used to restrict press freedom, right to information access, freedom of expression, and freedom of assembly and association.
- Seventh, new technologies have created unprecedented new crimes such as sexual violence through deepfakes, sexual exploitation, online bullying, and financial information theft. The UN Special Rapporteur on the Right to Privacy has also noted that new technologies have diversified and amplified forms of gender-based violence.

Theoretical and Practical Gaps in Responding to Various Human Rights Issues which Can Be Raised Through Digital New Technologies under the Current International Human Rights System

- The first type of gap is conceptual. Until now, international human rights norms have been created to address and resolve human rights violation issues based on offline rather than online situations, so the realities of the digital age have not been sufficiently reflected. Of course, this does not mean that the UN takes the position that adopting new human rights treaties or international agreements or amending existing documents is the best way to solve new technology and human rights issues.
- Gaps in expertise. Those working in new technology fields generally lack understanding of human rights, and conversely, human rights experts inevitably lack understanding of technology. This can create paradoxical situations of so-called human rights trade-offs where new technologies enhance the enjoyment of certain human rights while simultaneously violating other human rights. Without human rights-based established guidelines or norms, new technology system designers and business practitioners may choose and protect only specific human rights that are convenient for them. As a result, self-regulatory ethical codes of conduct dealing with a limited range of human rights tailored to individual companies' preferences can become prevalent in society, and we can actually confirm such movements today.

Theoretical and Practical Gaps in Responding to Various Human Rights Issues which Can Be Raised Through Digital New Technologies under the Current International Human Rights System

- Second, operational gaps can be considered. There is an inevitable time gap between the development of new technologies and the norms to regulate them. Social consensus is required first for the creation of new norms. Therefore, governments have no choice but to hope that the private sector will voluntarily comply with existing human rights norms, at least temporarily until norm creation. Such operational time gaps also cause various problems at the international governance level. The influence and impact of new technologies are global and transnational, but regulations on new technologies have so far been limited to discussions at national or regional levels. In practice, when international organizations and individual UN member states independently implement new technology-related policies, overlapping regulations and operations between these policies are likely to become inevitable. Therefore, international consultations and discussions between individual countries and international organizations are essential to avoid such overlaps and narrow operational gaps.

Theoretical and Practical Gaps in Responding to Various Human Rights Issues which Can Be Raised Through Digital New Technologies under the Current International Human Rights System

- Meanwhile, as the role of the private sector in protecting human rights increases, operational gaps may arise in relations with governments. While the primary obligation for human rights protection and promotion still lies with states, discussions on human rights protection obligations and roles that the private sector should bear have made considerable progress over the past decade since the adoption of the UN Guiding Principles on Business and Human Rights (UNGPs) in 2011. However, some innovative business models based on new technologies exist in ambiguous areas of law or are sometimes designed to exploit such areas, making it necessary to carefully consider what roles and obligations the private sector, including tech companies, should bear.

22

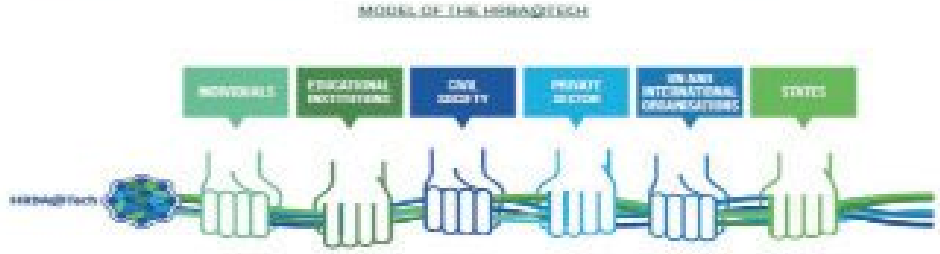
Comprehensive and Integrated Human Rights-Based Approach to New Technologies

- First, comprehensive understanding of technology is needed. This means not only understanding individual technologies, but also a comprehensive understanding of the unique complexity and interdependence of new technologies and datafication processes. The interconnectedness of various types of technological development and innovation-related human rights issues must be sufficiently grasped. It is also important to understand potential human rights-related issues that may arise throughout the technology development process from design, implementation, execution to disposal stages.
- Second, there must be a comprehensive approach to human rights. This requires explaining and expressing individual human rights norms in common language that companies and technology developers can practically understand. If the first element means that human rights experts should strive to understand new technologies, the second element conversely means that technology developers need to understand the language and content of human rights. This is even more so because many key decisions that will affect human rights are made by developers and engineers from the early stages of technology development. Therefore, from the early stages of new technology design and development, consideration for protecting and promoting basic human rights of civil rights, social rights, and social minorities such as women, persons with disabilities, and children should be included.

23

Comprehensive and Integrated Human Rights-Based Approach to New Technologies

- Third, comprehensive efforts for new technology and human rights governance and regulation by individual countries, international organizations and all relevant stakeholders are needed. The impact of new technologies on human rights can only be effectively regulated through cooperation among multiple stakeholders. Therefore, to effectively realize and strengthen human rights protection systems for new technologies and simultaneously establish sustainable monitoring systems, meaningful participation of various actors such as the private sector, academia, and especially civil society must be ensured.



24

UN Human Rights Council resolutions (2021)

United Nations A/HRC/RES/47/23

General Assembly

Distr.: General
14 July 2021

Original: English

Human Rights Council
Forty-seventh session
21 June–12 July 2021
Agenda item 3
Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

Resolution adopted by the Human Rights Council on 13 July 2021

47/23. New and emerging digital technologies and human rights

The Human Rights Council,

Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

3. *Requests* the Office of the High Commissioner to convene two expert consultations, to discuss the relationship between human rights and technical standard-setting processes for new and emerging digital technologies and the practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, and to submit a report thereon, reflecting the discussions held in an inclusive and comprehensive manner, to the Human Rights Council at its fiftieth and fifty-third sessions;

4. *Also requests* the Office of the High Commissioner, when preparing the above-mentioned expert consultations and reports, to seek input from and to take into account the relevant work already done by stakeholders from diverse geographic regions, including States, international and regional organizations, the Advisory Committee, the special procedures of the Human Rights Council, the treaty bodies, other relevant United Nations agencies, funds and programmes, including the International Telecommunication Union, other relevant standard development organizations, and the Office of the Envoy of Secretary-General on Technology, within their respective mandates, national human rights institutions, civil society, the private sector, the technical community and academic institutions;

25

Linkage with Business and Human Rights Discussions

- UN discussions on new technologies and human rights have mainly been conducted in connection with business and human rights discussions for soft law international norm-making.
- In particular, on July 30, 2019, the UN Human Rights Council Working Group on Business and Human Rights launched the B-Tech Project on digital technology, raising the need to respond to negative impacts on human rights arising from the development of digital technology. The main purpose of the project is to prepare practical information, guidance and recommendations for states and companies through comprehensive consultation and research with the participation of various companies, civil society organizations and policy stakeholders. The B-Tech Project is specifically examining the following four strategic focus areas:
  - Business models that can respond to human rights risks
  - Human rights due diligence and end-use
  - Accountability and remedy
  - Smart mix of measures

UNITED NATIONS HUMAN RIGHTS

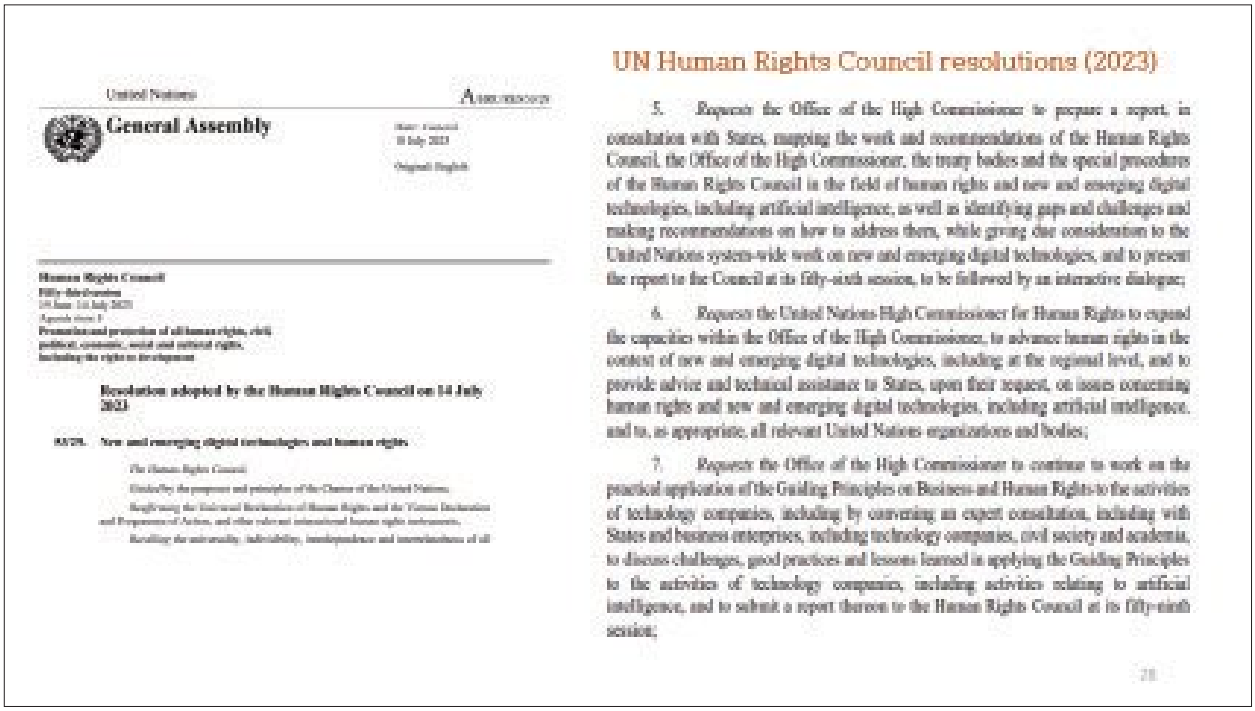
B-Tech

26

Linkage with Business and Human Rights Discussions

- In April 2022, "The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of Technology Companies" (A/HRC/50/56, April 21, 2022) was published, presenting more specific implementation measures.
- The report emphasizes that the state's duty to protect, corporate responsibility to respect, and access to remedy mentioned in the "Guiding Principles on Business and Human Rights" should all be applied in the technology industry, and particularly actively encourages companies in this field to conduct human rights due diligence in the process of designing, developing and using new technologies. It also emphasizes the need to pay more attention to vulnerable groups considering digital divide and gender equality.

27



United Nations A/HRC/RES/53/29

General Assembly

Distr.: General
14 July 2023

Original: English

Human Rights Council
Fifty-third session
19 June–14 July 2023
Agenda item 3
Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

Resolution adopted by the Human Rights Council on 14 July 2023

53/29. New and emerging digital technologies and human rights

The Human Rights Council,

Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

Recalling the universality, indivisibility, interdependence and interrelatedness of all

UN Human Rights Council resolutions (2023)

5. *Requests* the Office of the High Commissioner to prepare a report, in consultation with States, mapping the work and recommendations of the Human Rights Council, the Office of the High Commissioner, the treaty bodies and the special procedures of the Human Rights Council in the field of human rights and new and emerging digital technologies, including artificial intelligence, as well as identifying gaps and challenges and making recommendations on how to address them, while giving due consideration to the United Nations system-wide work on new and emerging digital technologies, and to present the report to the Council at its fifty-sixth session, to be followed by an interactive dialogue;

6. *Requests* the United Nations High Commissioner for Human Rights to expand the capacities within the Office of the High Commissioner, to advance human rights in the context of new and emerging digital technologies, including at the regional level, and to provide advice and technical assistance to States, upon their request, on issues concerning human rights and new and emerging digital technologies, including artificial intelligence, and to, as appropriate, all relevant United Nations organizations and bodies;

7. *Requests* the Office of the High Commissioner to continue to work on the practical application of the Guiding Principles on Business and Human Rights to the activities of technology companies, including by convening an expert consultation, including with States and business enterprises, including technology companies, civil society and academia, to discuss challenges, good practices and lessons learned in applying the Guiding Principles to the activities of technology companies, including activities relating to artificial intelligence, and to submit a report thereon to the Human Rights Council at its fifty-sixth session;

28

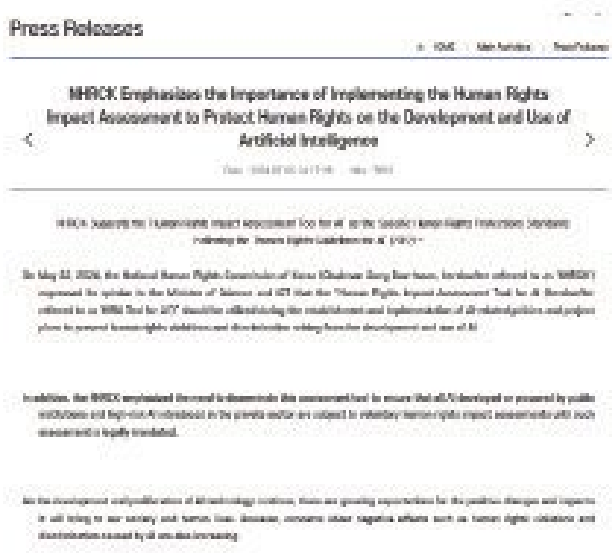
Artificial Intelligence and Human Rights Related Discussions

- Looking at the characteristics of international society discussions so far, legal issues surrounding the operation and regulation of AI systems are being discussed almost entirely based on the existing international human rights law system. However, there is no clear answer as to whether the existing legal system can sufficiently encompass all issues surrounding AI ethics. The scope of international human rights law cited extends to international treaties as well as declarations, resolutions, recommendations or guidelines, regardless of the legal effect of the cited international documents. Moreover, individual countries' technical, financial capabilities and social environments vary greatly, so gaps are inevitable in the process of domestically implementing various recommendations. In specific implementation processes, cherry-picking or modification phenomena by individual countries may occur, and if so, AI ethics-related recommendations and guidelines may result in unequal, fragmented or diluted outcomes instead of achieving their intended goals. This shows the limitations of the voluntary implementation process of non-binding recommendations and guidelines.

29

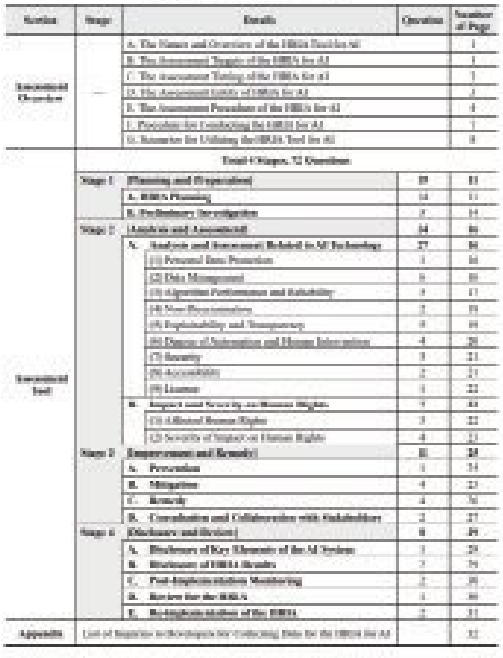
Artificial Intelligence and Human Rights Related Discussions

- There is unanimous agreement that it is necessary to prevent artificial intelligence deeply embedded in our daily lives from violating human rights and to establish systems that can respond to human rights violations when they occur. However, systems to coordinate when users' right to know conflicts with AI developers' rights to protect trade secrets are needed. For example, pharmaceuticals traditionally adopt a licensing system that diagnoses the risks of new drugs before product launch and allows them to be launched in the market only when determined safe for general public use. Similarly, for AI software, an AI Software Preecertification Program can be considered that tests software in the development stage and allows modifications during development if problems are found, rather than inspecting software that has already been developed and launched. Also, since negative impacts caused by new technologies are extensive, human rights impact assessments must be conducted as early as possible to prevent potential impacts. Furthermore, not only companies that plan and develop new technologies, but also companies in other industries (sector-agnostic) that introduce and utilize specific new technologies must mandatorily implement human rights impact assessments.



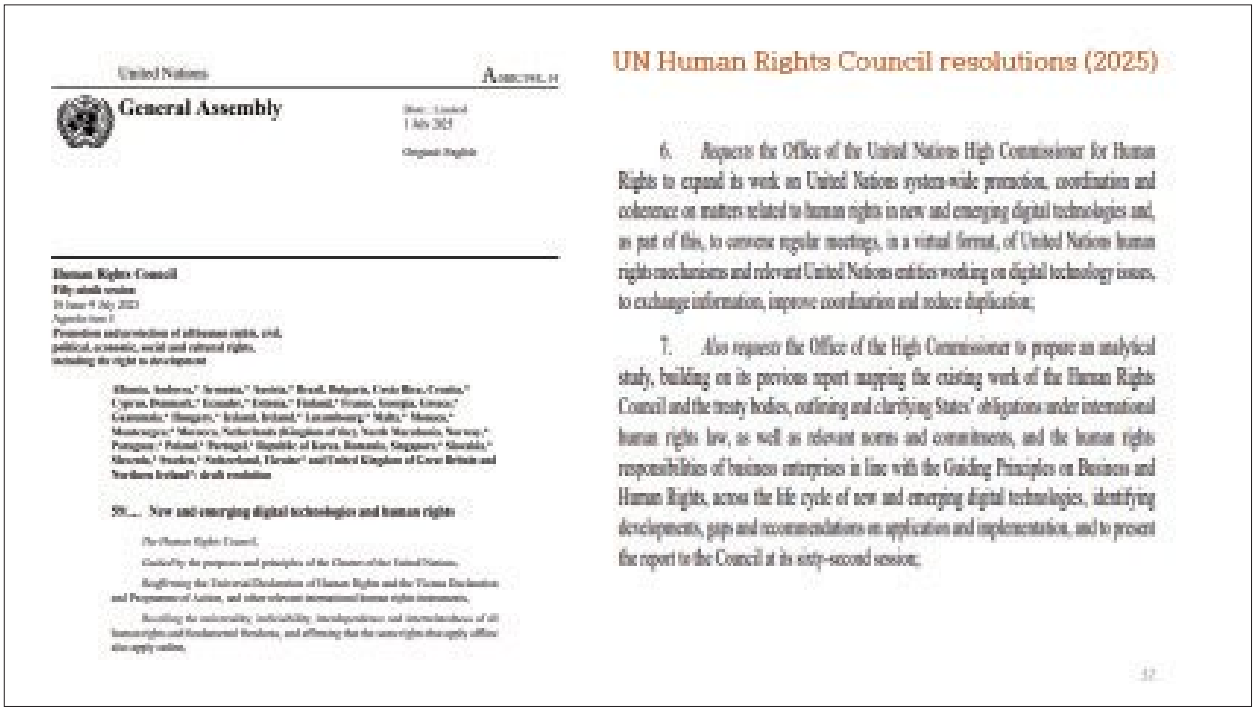
Press Releases

NHRCK Emphasizes the Importance of Implementing the Human Rights Impact Assessment to Protect Human Rights on the Development and Use of Artificial Intelligence



Components of an Assessment Tool

Section	Stage	Details	Question	Number of Page
Assessment Overview	—	A. The Nature and Overview of the HRIA Tool for AI		1
		B. The Assessment Targets of the HRIA for AI		1
		C. The Assessment Timing of the HRIA for AI		1
		D. The Assessment Units of HRIA for AI		1
		E. The Assessment Procedure of the HRIA for AI		1
		F. Principles for Conducting the HRIA for AI		1
		G. Scenarios for Utilizing the HRIA Tool for AI		1
Assessment Tool	Total 4 Stages, 72 Questions			
	Stage 1	[Planning and Preparation]	19	11
		A. HRIA Planning	14	11
		B. Preliminary Investigation	5	11
	Stage 2	[Analysis and Assessment]	34	16
		A. Analysis and Assessment Related to AI Technology	27	16
		(1) Personal Data Protection	1	16
		(2) Data Management	1	16
		(3) Algorithm Performance and Reliability	3	17
		(4) Non-Discrimination	2	18
		(5) Explainability and Transparency	3	19
		(6) Degree of Automation and Human Intervention	4	20
		(7) Security	3	21
		(8) Accountability	2	21
		(9) Licensee	1	22
		B. Impact and Severity on Human Rights	7	22
		(1) Affected Human Rights	3	22
		(2) Severity of Impact on Human Rights	4	23
	Stage 3	[Improvement and Remedy]	11	24
		A. Prevention	3	24
		B. Mitigation	4	25
		C. Remedy	4	26
		D. Consultation and Collaboration with Stakeholders	2	27
	Stage 4	[Disclosure and Review]	8	28
		A. Disclosure of Key Elements of the AI System	3	28
		B. Disclosure of HRIA Results	3	29
		C. Post-Implementation Monitoring	2	30
		D. Review for the HRIA	1	30
		E. Re-implementation of the HRIA	1	31
Appendix	List of Inquiries to Developers for Collecting Data for the HRIA for AI			32



United Nations A/HRC/59/L.11

General Assembly

Distr.: Limited
1 July 2025

Original: English

Human Rights Council
Fifty-ninth session
16 June–9 July 2025
Agenda item 3
Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

Albania, Andorra,* Armenia,* Austria,* Brazil, Bulgaria, Costa Rica, Croatia,* Cyprus, Denmark,* Ecuador,* Estonia,* Finland,* France, Georgia, Greece,* Guatemala,* Hungary,* Iceland, Ireland,* Luxembourg,* Malta,* Monaco,* Montenegro,* Morocco, Netherlands (Kingdom of the), North Macedonia, Norway,* Paraguay,* Poland,* Portugal,* Republic of Korea, Romania, Singapore,* Slovakia,* Slovenia,* Sweden,* Switzerland, Ukraine* and United Kingdom of Great Britain and Northern Ireland*: draft resolution

59/... New and emerging digital technologies and human rights

The Human Rights Council,

Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

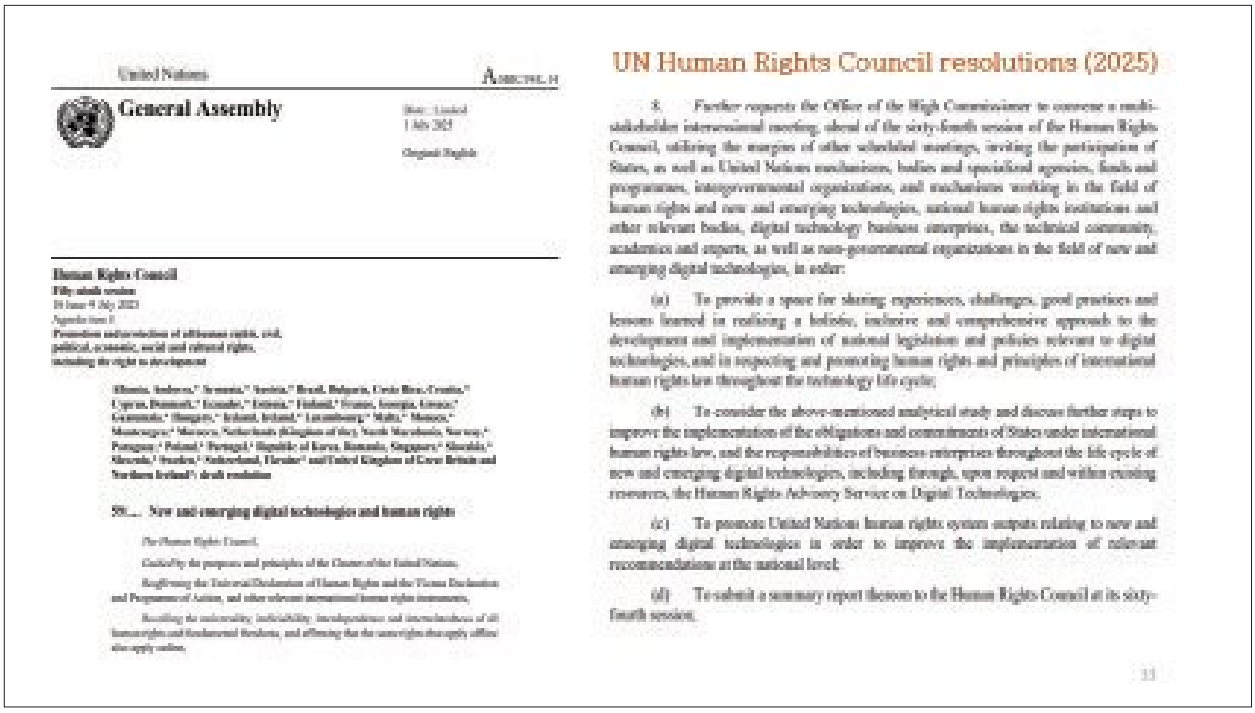
Recalling the universality, indivisibility, interdependence and interrelatedness of all human rights and fundamental freedoms, and affirming that the same rights also apply offline also apply online,

UN Human Rights Council resolutions (2025)

6. Requests the Office of the United Nations High Commissioner for Human Rights to expand its work on United Nations system-wide promotion, coordination and coherence on matters related to human rights in new and emerging digital technologies and, as part of this, to convene regular meetings, in a virtual format, of United Nations human rights mechanisms and relevant United Nations entities working on digital technology issues, to exchange information, improve coordination and reduce duplication;

7. Also requests the Office of the High Commissioner to prepare an analytical study, building on its previous report mapping the existing work of the Human Rights Council and the treaty bodies, outlining and clarifying States' obligations under international human rights law, as well as relevant norms and commitments, and the human rights responsibilities of business enterprises in line with the Guiding Principles on Business and Human Rights, across the life cycle of new and emerging digital technologies, identifying developments, gaps and recommendations on application and implementation, and to present the report to the Council at its sixty-second session;

12



United Nations A/HRC/59/L.11

General Assembly

Distr.: Limited
1 July 2025

Original: English

Human Rights Council
Fifty-ninth session
16 June–9 July 2025
Agenda item 3
Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development

Albania, Andorra,* Armenia,* Austria,* Brazil, Bulgaria, Costa Rica, Croatia,* Cyprus, Denmark,* Ecuador,* Estonia,* Finland,* France, Georgia, Greece,* Guatemala,* Hungary,* Iceland, Ireland,* Luxembourg,* Malta,* Monaco,* Montenegro,* Morocco, Netherlands (Kingdom of the), North Macedonia, Norway,* Paraguay,* Poland,* Portugal,* Republic of Korea, Romania, Singapore,* Slovakia,* Slovenia,* Sweden,* Switzerland, Ukraine* and United Kingdom of Great Britain and Northern Ireland*: draft resolution

59/... New and emerging digital technologies and human rights

The Human Rights Council,

Guided by the purposes and principles of the Charter of the United Nations,

Reaffirming the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action, and other relevant international human rights instruments,

Recalling the universality, indivisibility, interdependence and interrelatedness of all human rights and fundamental freedoms, and affirming that the same rights also apply offline also apply online,

UN Human Rights Council resolutions (2025)

8. *Further requests* the Office of the High Commissioner to convene a multi-stakeholder intersessional meeting, ahead of the sixty-fourth session of the Human Rights Council, utilizing the margins of other scheduled meetings, inviting the participation of States, as well as United Nations mechanisms, bodies and specialized agencies, funds and programmes, intergovernmental organizations, and mechanisms working in the field of human rights and new and emerging technologies, national human rights institutions and other relevant bodies, digital technology business enterprises, the technical community, academics and experts, as well as non-governmental organizations in the field of new and emerging digital technologies, in order:

(a) To provide a space for sharing experiences, challenges, good practices and lessons learned in realizing a holistic, inclusive and comprehensive approach to the development and implementation of national legislation and policies relevant to digital technologies, and in respecting and promoting human rights and principles of international human rights law throughout the technology life cycle;

(b) To consider the above-mentioned analytical study and discuss further steps to improve the implementation of the obligations and commitments of States under international human rights law, and the responsibilities of business enterprises throughout the life cycle of new and emerging digital technologies, including through, upon request and within existing resources, the Human Rights Advisory Service on Digital Technologies;

(c) To promote United Nations human rights system outputs relating to new and emerging digital technologies in order to improve the implementation of relevant recommendations at the national level;

(d) To submit a summary report thereon to the Human Rights Council at its sixty-fourth session;

13

UN Discussion on Digital New Technologies and Human Rights

- The UN Human Rights Council has requested the Advisory Committee to prepare a report examining the human rights implications of emerging technologies in various fields.

Current mandates

- Human rights implications of new and emerging technologies in the military domain
- Impact of disinformation on the enjoyment and realization of human rights
- Technology-facilitated gender-based violence
- Implications of plastic pollution for the full enjoyment of human rights
- Impact of artificial intelligence systems on good governance
- Negative impact of unilateral coercive measures on the right to health
- Social justice through the domestic legal enforcement of economic, social and cultural rights
- Human rights guidelines on neurotechnology
- Negative impact of corruption on the enjoyment of human rights

Past mandates and achievements

- Neurotechnology and human rights
- Impact of new technologies for climate protection on the enjoyment of human rights
- Advancement of racial justice and equality
- New and emerging digital technologies and human rights

14

[발표 2 | Speaker 2]

신뢰할 수 있는 AI와 기본권 평가의 교훈

Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment

로베르토 지커리
Roberto V. ZICARI

서울대학교 데이터사이언스대학원 특임교수
Adjunct Professor, Graduate School of Data Science, Seoul National University

신뢰할 수 있는 AI와 기본권 평가의 교훈

로베르토 지커리 | 서울대학교 데이터사이언스대학원 특임교수

기술은 진공 상태에서 만들어지지 않습니다.

- "그것은 이를 설계한 사람들의 가치관, 관점, 그리고 권력 구조를 담고 있습니다."

출처: UN 여성 회의, 서울: https://sites.google.com/view/ai-and-gender-conference-websi/home

저희는 Z-Inspection® 이니셔티브입니다 (2019년 1월 시작)

Z-Inspection® 이니셔티브는 *비영리적* 이니셔티브입니다

100명 이상의 전문가

78개 *협력* 기관 및 연구소

전 세계 40개국에서 활동 중입니다.

호주, 오스트리아, 벨기에, 브라질, 캐나다, 칠레, 키프로스, 크로아티아, 덴마크, 에스토니아, 핀란드, 프랑스, 독일, 그리스, 헝가리, 아이슬란드, 인도, 아일랜드, 이탈리아, 일본, 라트비아, 리투아니아, 룩셈부르크, 말레이시아, 네덜란드, 나이지리아, 노르웨이, 폴란드, 포르투갈, 루마니아, 시에라리온, 대한민국, 스페인, 스웨덴, 스위스, 우간다, 터키, 영국, 미국, 뉴질랜드.

https://z-inspection.org

우리의 사명

Z-Inspection®을 통해 우리는
'AI의 신중한 사용(#MUAI)'을
정착시키는 것을 목표로 합니다.

본 연구의 동기

신뢰할 수 있는 AI를 실제로
어떻게 평가할 것인가?

사진 RVZ

Z-inspection® 프로세스

우리는 숙련된 전문가 팀이 특정 *맥락에서* AI 제품/서비스 사용의 ***윤리적, 기술적, 분야별*** 및 ***법적*** 함의를 평가할 수 있도록 지원하는 *참여형 프로세스*를 구축했습니다.

- IEEE Transactions on Technology and Society에 게재됨
  VOL. 2, NO. 2, 2021년 6월

"책임 있는 AI 활용" 시범 프로젝트

- ***프리스란 주, Rijks ICT Gilde 및 Z-Inspection® 이니셔티브와의 시범 프로젝트.***
- Marjolein Boonstra, Frédérick Bruneault, Subrata Chakraborty, Tjitske Faber, Alessio Gallucci, Eleanore Hickman, Gerard Kema, Heejin Kim, Jaap Kooiker, Elisabeth Hildt, Annegret Lamade, Emilie Wiinblad Mathez, Florian Möslein, Genien Pathuis, Giovanni Sartor, Marijke Steege, Alice Stocco, Willy Tadema, Jarno Tuimala, Isabel van Vledder, Dennis Vetter, Jana Vetter, Magnus Westerlund, Roberto V. Zicari.

- 본 시범 사업은 2022년 5월부터 2023년 1월까지 진행되었습니다.
- 시범 사업 기간 동안 프리슬란(Fryslân) 주에서 개발한 딥러닝 알고리즘을 실제로 적용하고 평가했습니다.

환경 모니터링

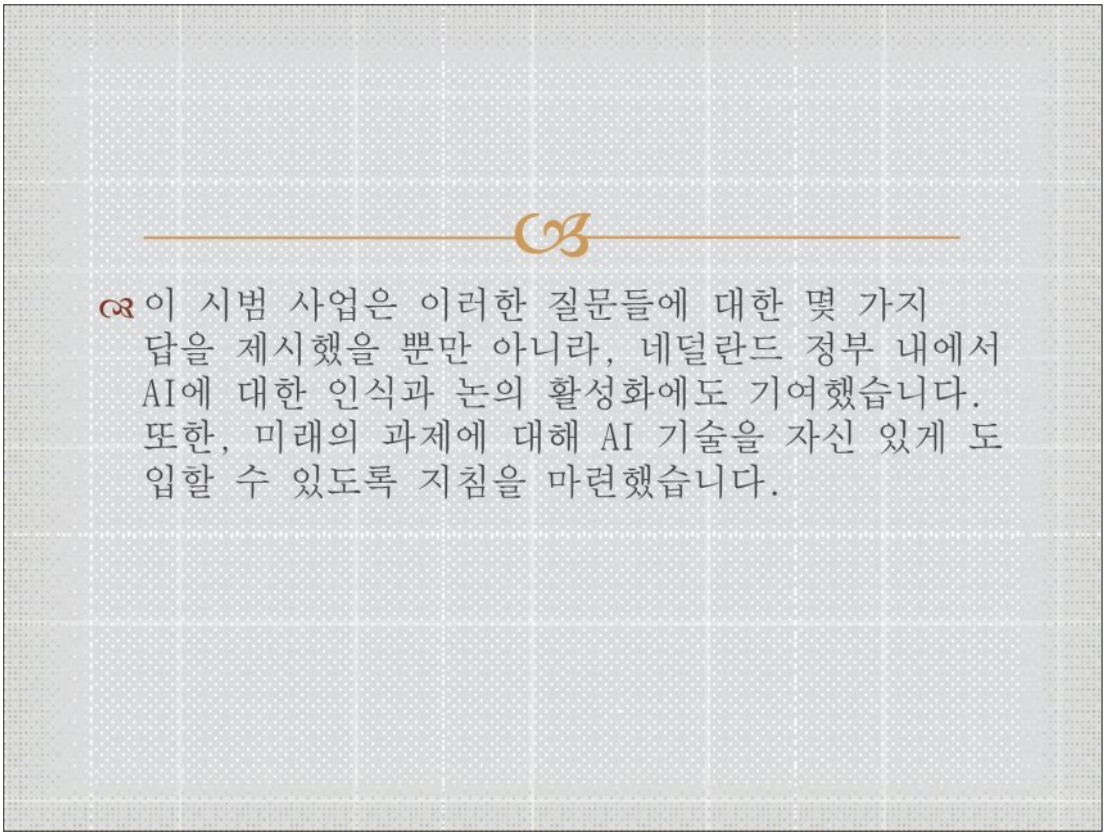
- AI는 자연 보호구역 모니터링을 위해 위성 이미지를 활용해 관목지대와 초지를 지도화 하였습니다. **환경 모니터링은** 식수 기준을 유지하는 것부터 특정 국가나 지역의 이산화탄소 배출량 측정에 이르기까지 다양한 목적을 위해 이루어지는 중요한 활동입니다.
- 위성 영상과 머신러닝을 활용해 의사 결정을 지원하는 것은 환경 모니터링에서 점점 중요한 부분이 되고 있다.

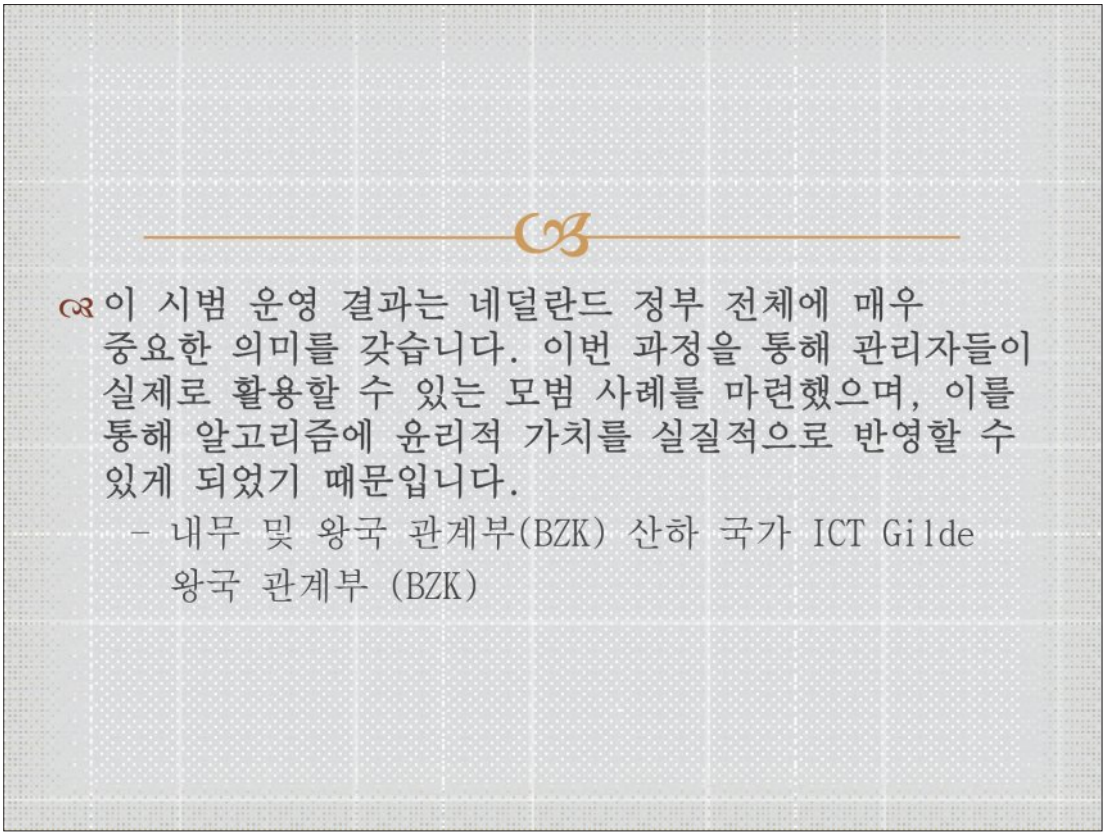
실행 과정에서 얻은 경험, 결과 및 교훈의 공유

- **신뢰할 수 있는 AI 평가를** 위한 Z-Inspection® 프로세스와 EU 신뢰할 수 있는 AI 프레임워크를 활용하여 신뢰할 수 있는 AI 평가를 수행하였으며,
- 네덜란드 정부가 공공기관의 AI 알고리즘 활용 시 권장하는 기본권 및 알고리즘 영향평가 (FRAIA)를 활용한 **기본권 평가**도 함께 진행했습니다.

"신뢰할 수 있는 AI 평가" 시범 사업은 다음과 같은 질문에 답하고자 했습니다

- 정부로서 *책임 있는* AI의 개발과 활용을 어떻게 관리할 수 있는가요?
- AI 개발과 활용에서 어떤 프레임워크, 법률 및 규정이 중요하며, 이를 어떻게 평가합니까?
- AI 응용 프로그램을 어떻게 분석하고 평가하며 개선합니까?
- 그리고 이러한 응용 프로그램이 공공 가치와 인권에 부합하나요?
- AI 시스템이 제기하는 윤리적 문제는 무엇인가요?
- AI 시스템으로 인해 어떤 기본권을 영향을 받을 수 있나요?
- AI 시스템이 *신뢰할* 수 있도록 하기 위해 어떤 조치를 취할 수 있을까요?


이 시범 사업은 이러한 질문들에 대한 몇 가지 답을 제시했을 뿐만 아니라, 네덜란드 정부 내에서 AI에 대한 인식과 논의 활성화에도 기여했습니다. 또한, 미래의 과제에 대해 AI 기술을 자신 있게 도입할 수 있도록 지침을 마련했습니다.


이 시범 운영 결과는 네덜란드 정부 전체에 매우 중요한 의미를 갖습니다. 이번 과정을 통해 관리자들이 실제로 활용할 수 있는 모범 사례를 마련했으며, 이를 통해 알고리즘에 윤리적 가치를 실질적으로 반영할 수 있게 되었기 때문입니다.
- 내무 및 왕국 관계부(BZK) 산하 국가 ICT Gilde
왕국 관계부 (BZK)

배경

- 프리슬란(Fryslân) 주는 앞으로 데이터의 스마트하고 효과적인 활용을 위해 집중적으로 투자할 계획입니다. 주 정부는 거의 모든 지역 개발과 사회적 과제에 데이터가 관여하고 있음을 인식하고 있습니다. 이에 따라 데이터와 AI에 대한 관심과 대응의 필요성이 어느 때 보다 커지고 있습니다. 기술 발전에 책임 있게 대응하기 위해서는 데이터와 AI에 대한 명확한 비전이 필요합니다. 이번 파일럿 프로젝트 참여는 미래의 디지털 인프라를 설계하고, 이를 뒷받침할 윤리적 프레임워크를 수립하는데 큰 도움이 되었습니다.

- 이 시범 사업과 직접적으로 관련하여, 프리슬란 주(Fryslân)는 자연 보호 구역의 생물 다양성을 모니터링하도록 법적으로 규정되어 있습니다. 현재는 10년에 한 번씩 수작업으로 시각적 모니터링을 통해 이를 수행하고 있습니다. 하지만 자연 보호 구역을 더 자주 모니터링 하고 지도화하며, 특히 관목 지대의 초지화 현상을 보다 신속하게 파악할 필요가 있습니다. 이 작업을 원활하게 하고 비용을 줄이며 절차를 간소화하기 위해, 프리슬란 주는 AI 시스템 개발을 제3자에게 의뢰했습니다

시범 사업의 범위

- 본 시범 사업의 범위는 이 AI 시스템이 신뢰할 만한지,어떤 기본적 인권에 영향을 미치는지, 그리고 이를 책임 있게 활용하기 위해 무엇이 필요한지를 평가하는 것이었습니다.

AI 시스템의 목적

- 이 AI 시스템의 목적은 생태학자들이 **자연 지역을 신속하고 자주 관측**해, 설정된 **자연 환경의 질 목표가 달성되고 있는지, 적절한 관리 조치가 이루어지고 있는지**, 그리고 **생물 다양성 증진을 위한 접근 방식이 효과적인지**를 확인할 수 있도록 돕는 것입니다.
- 특히, 이 AI 시스템은 **위성 이미지**를 활용해, 관목 지대에서 **침입성이 강하고 원치 않는 잡초인** Molinia caerulea과 Avenella flexuosa**의 확산에 관한 정보를 제공하는 것을 목표로 합니다.** 사용되는 위성 이미지는 네덜란드 우주청(NSO)에서 제공합니다. 이는 GIS에서 활용할 수 있는 고해상도 광학 위성 이미지로, 위성 데이터 포털을 통해 무료로 이용할 수 있습니다.

접근법

본 시범 사업에서는 AI 시스템을 세 가지 관점에서 검토하였습니다:

- 1. 기술적
- 2. 생태학적
- 3. 윤리와 기본권

윤리 및 기본권 평가

2022년 3월 네덜란드에서 알고리즘에 대한 기본권 영향 평가 도구가 도입됨에 따라, 이번 시범 사업에서는 혼합형 접근법이 채택했습니다.

- 먼저, **FRAIA를** 활용하여 AI 시스템이 **인권** 기준에 따라 평가했습니다. 이 과정에서는 인권 침해 가능성뿐 아니라, AI 시스템을 통해 보호되거나 강화될 수 있는 권리까지 고려했습니다. 예를 들어 건강한 환경에 대한 권리가 이에 포함됩니다.
- 그 후, 유럽연합의 신뢰할 수 있는 AI 가이드라인에 따라 AI 시스템이 제기하는 윤리적 쟁점들을 식별하고 평가했으며, 이 시스템을 보다 넓은 관점에서 종합적으로 검토했습니다.

주장, 논거 및 근거

우리의 접근법은 '주장, 논거, 근거(*Claim, Arguments and Evidence*)' 프레임워크를 활용해 기본권 평가와 신뢰할 수 있는 AI 평가를 결합한 근거 기반 접근법이라는 점에서 독창적입니다.

포커스: 기본권

- AI 시스템에 의해 영향을 받는 기본권을 식별하기 위해, 이번 평가에서는 FRAIA가 제시한 기본권 목록을 활용했습니다. 이 목록을 네 개의 그룹으로 나뉘며, 각 그룹 아래에 세부 권리가 포함되어 있습니다.
- 개인의 권리
- 자유와 관련된 기본권
- 평등권
- 절차적 기본권

포커스: 윤리

- 이어서, 평가에서는 AI 시스템에서 발생할 수 있는 윤리적 쟁점을 보다 넓은 관점에서 검토했습니다. 특히, EU AI 고위 전문가 그룹 (AI HLEG, 2019)이 정의한 신뢰할 수 있는 인공지능을 위한 윤리 가이드라인을 참고했습니다. 이 가이드라인은 상호 간에 긴장이 발생할 수 있음을 인정하면서, 다음 4가지 윤리 원칙을 제시하고 있습니다.
- (1) 인간의 자율성 존중 (2) 피해 예방 (3) 공정성 (4) 설명 가능성
- 또한 (AI HLEG, 2019)에서 정의한 신뢰할 수 있는 AI의 7가지 요구사항이 검토했으며, 각 요구사항은 여러 하위 요구사항으로 구성되어 있습니다.

EU 신뢰할 수 있는 인공지능 프레임워크의 4대 윤리 원칙

기본권에 기반한 4대 윤리 원칙

(i) 인간 자율성 존중
(ii) 피해 예방
(iii) 공정성
(iv) 설명 가능성

- 이러한 원칙들 간에 **상충 관계가** 발생할 수 있습니다.

출처: *신뢰할 수 있는 인공지능을 위한 윤리 지침*. 인공지능에 관한 독립적 고위 전문가 그룹. 유럽 위원회. 2019년 4월 8일.

23

EU 신뢰할 수 있는 AI를 위한 7대 요구사항 및 세부 항목

24

출처: *신뢰할 수 있는 인공지능을 위한 윤리 지침*. 인공지능에 관한 독립 고위 전문가 그룹. 유럽 위원회, 2019년 4월 8일.

- 요구사항 및 세부 항목
- 1 인간의 주체성 및 감독(Human agency and oversight) *기본권, 인간의 주체성, 인간의 감독 포함*
- 2 기술적 견고성 및 안전성(Technical robustness and safety) *공격에 대한 복원력과 보안, 비상 대응 계획 및 일반적 안전성, 정확성, 신뢰성, 재현성 포함*
- 3 프라이버시와 데이터 거버넌스(Privacy and data governance) *프라이버시 존중, 데이터의 품질과 무결성, 데이터 접근성 포함*
- 4 투명성(Transparency) *추적 가능성, 설명 가능성, 의사소통 포함*
- 5 다양성, 비차별, 공정성(Diversity, non-discrimination and fairness) *불공정한 편향의 회피, 접근성 및 보편적 설계, 이해관계자 참여 포함*
- 6 사회적·환경적 복지(Societal and environmental well-being) *지속 가능성, 친환경성, 사회적 영향, 사회와 민주주의 포함*
- 7 책임성(Accountability) *감사 가능성, 부정적 영향의 최소화 및 보고, 상충 관계 관리, 구제 절차 포함*

증거 기반 접근법을 통한 인권 평가

이 시범 사업에서는 영향을 받을 가능성이 있는 권리 각각에 대해, 평가를 통해 다음 중 하나의 결론을 내렸습니다.

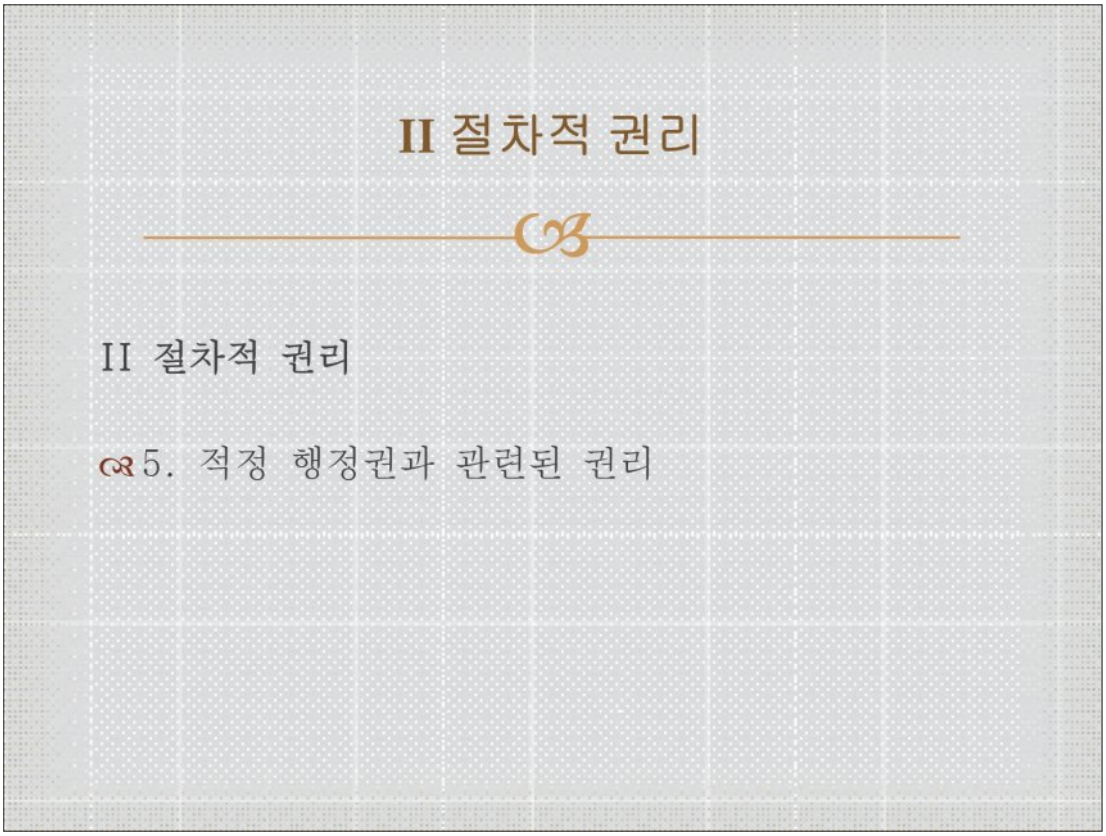
- a) 영향을 받음(긍정적이든 부정적이든 관계없이)
- b) 영향을 받지 않음
- c) 특정 조건이나 추가 설명에 따라 영향을 받을 가능성이 있음

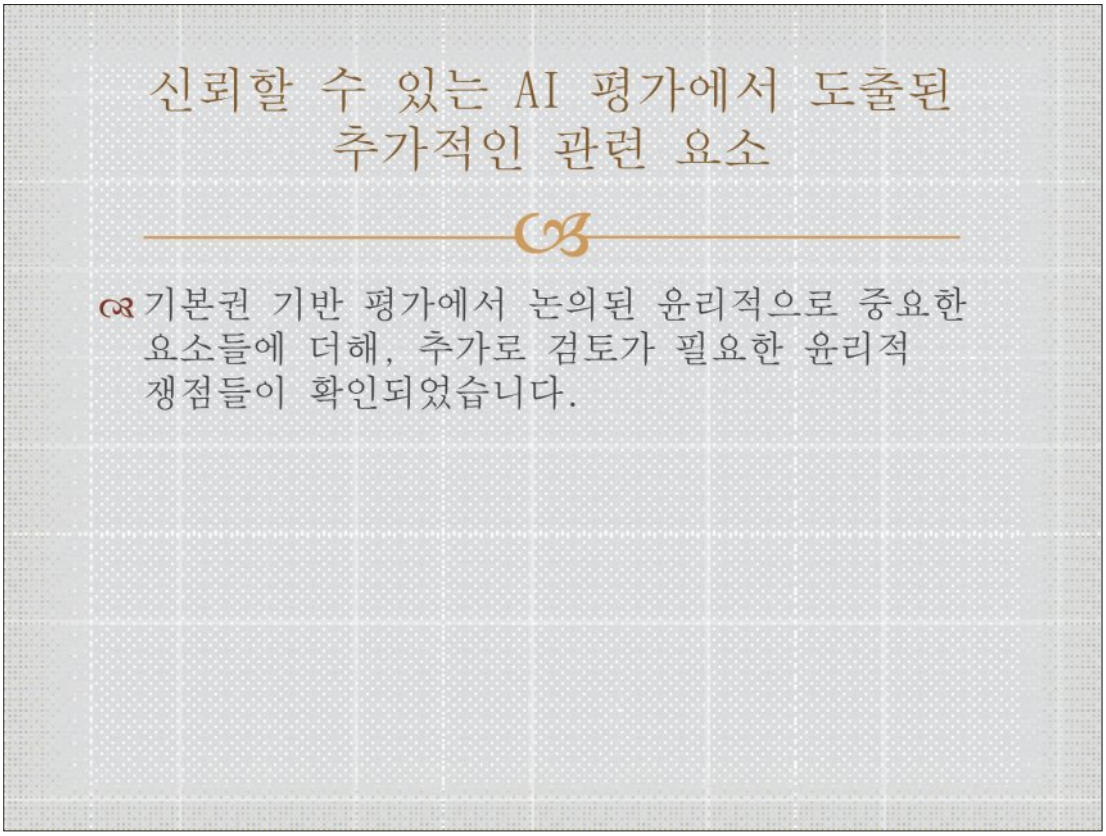
각 주장에 대해 간단한 논거를 제시되고, 해당 권리가 영향을 받는지를 뒷받침하는 근거가 함께 제공합니다. 평가 결과, 이 AI 시스템에 의해 영향을 받을 가능성이 있는 5개의 기본권 그룹이 확인되었습니다.

I. 개인과 관련된 권리

I. 개인과 관련된 권리:

- 건강한 생활 환경에 관한 권리
- 개인 정체성, 인격권, 개인적 자율성과 관한 권리
- 데이터 보호 및 정보 프라이버시 권리
- 공간적 프라이버시와 관련된 권리


II 절차적 권리
II 절차적 권리
5. 적정 행정권과 관련된 권리


신뢰할 수 있는 AI 평가에서 도출된 추가적인 관련 요소
기본권 기반 평가에서 논의된 윤리적으로 중요한 요소들에 더해, 추가로 검토가 필요한 윤리적 쟁점들이 확인되었습니다.

윤리적 문제

- 투명성과 비투명성
- 관련 정보의 수신
- 인간의 주체성과 감독
- 기술적 견고성 및 안전성
- 정의와 공정성
- 비용 절감
- 다양성과 포용성
- 책임과 책무성
- 의사 결정 과정의 적정 절차(Due diligence)

신뢰할 수 있는 AI평가 과정과 기본권 기반 FRAIA 평가 도구의 비교

- FRAIA 도구는 유용하고 명확했지만, 여전히 인권 평가를 어떻게 구성할 것인가에 대한 질문이 제기되었습니다. 특히, AI 시스템의 신뢰성과 윤리적 측면을 평가하는 과정에서 기본권을 어떻게 고려해야 하는지가 핵심 쟁점이었습니다.
- 기본권을 **법률의 정의와 법원의 해석에 한정해** 바라봐야 할까요?
- 아니면 이를 **윤리적 원칙과 연결해 보다 넓은 관점에서** 평가에 포함해야 할까요?
- 만약 **법적 정의에만 의존**한다면, 특정 법률이 적용되는지를 별도로 평가해야 합니다.

이중적이고 통합된 접근법

- 인권 평가의 범위를 너무 좁게 설정하면 윤리 평가나 신뢰성 평가와 분리된 별개의 평가가 될 위험이 있습니다.
- 반대로 평가의 범위를 너무 넓게 설정하면, 인권 기준이 희석될 위험이 있습니다.

따라서 법적 요건과 더 넓은 윤리적 쟁점을 함께 살펴보는 이중적이고 통합된 접근법을 고려할 수 있으며, 이는 조직의 구조와 활용 사례에 따라 달라질 수 있습니다.

두 가지 평가 접근법에서 얻은 교훈

- 기본권 평가와 **신뢰할 수 있는 AI 가이드라인** (Trustworthy AI) 가이드라인에 기반한 윤리 평가는 서로 보완적인 관계에 있으며, 두 접근법 모두 AI 활용 사례에 대한 중요한 통찰을 제공합니다.

유사점과 차이점

윤리적 관점에서 AI를 살펴보는 것은 기본권 평가와 밀접하게 연결되어 있습니다.

- 윤리와 기본권 모두 사회가 지닌 규범과 근본적인 가치와 관련되어 있기 때문입니다.
- 또한, 윤리적 성찰과 가이드라인은 법에 영향을 미치기 때문에, 기술의 발전과 그 사회적 영향을 논의할 때는 두 분야의 학자들이 함께 협력해야 합니다.
- 두 접근법은 많은 공통점을 가지고 있지만, 동시에 여러가지 중요한 차이점도 존재합니다.

차이점

- **기본권 기반 접근법은 기존 법체계와 더 밀접하게 연결되어 있으며,법적으로 관련성이 있고 집행 가능한 측면에 초점을 맞춥니다.**
- **반면, 윤리 기반 접근법은 훨씬 더 범위가 넓고 개방적이며, 법적 관점에서 고려할 가치가 없을 수 있는 잠재적 영향까지도 성찰합니다.**
- 예를 들어, 이번 파일럿 프로젝트의 AI 도구와 관련해 윤리적 관점에서는 개인의 자율성, 의사결정의 자유, 공정성이 매우 중요한 개념임을 확인하였습니다. 하지만, 기본권 관점, 즉 엄격한 법적 의미에서의 개인 자율성 관련 권리는 이 AI 도구에 의해 침해되지 않은 것으로 판단되었습니다.

다양한 관점

- **기본권 기반 평가**는 기본권이 **부정적으로 영향을 받거나 침해되었는지**에 초점을 맞춥니다. 반면, **윤리적 관점**에서는 AI 기술이 미치는 **긍정적 영향과 부정적 영향**을 모두 고려합니다.
- 예를 들어, 이번 파일럿에서는 이 AI 도구가 **환경에 미칠 수 있는 잠재적인 긍정적 영향**이 핵심적인 요소로 확인되었습니다.

최종 참고 사항

- 기본권 관점에서 사례를 다룬다는 것은, AI의 **윤리적·사회적 측면과 그 영향**을 **기본권 및 기존 법률과의 관련성**이 있는 범위 내에서만 논의한다는 것을 의미합니다.

자세한 정보

신뢰할 수 있는 AI 및 기본권 평가 수행에서 얻은 교훈(Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment)
인용: arXiv:2404.14366 [cs.CY]
https://arxiv.org/abs/2404.14366

YouTube 동영상:
https://www.youtube.com/watch?v=z_RCyscIXdk

2025 ISSIP Awards *Excellence in Service Innovation*
Distinguished Recognition - Impact to Society 부문 수상
https://z-inspection.org/wp-content/uploads/2025/03/ISSIP2025-certificate.pdf

Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment

Roberto V. ZICARI | Adjunct Professor, Graduate School of Data Science, Seoul National University

Technology is not created in a vacuum.

- **"It carries the values, perspectives, and power structures of those who design it."**

Source UN Women Conference, Seoul :https://sites.google.com/view/ai-and-gender-conference-websi/home

We are the Z-Inspection® Initiative (started Jan. 2019)

The Z-Inspection® Initiative is a *non-commercial* initiative

100+ individual experts

78 *affiliated* Institutions and Labs

in 40 countries all over the world.

Australia, Austria, Belgium, Brasil, Canada, Chile, Cyprus, Croatia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, India, Ireland, Italy, Japan, Latvia, Lithuania, Luxemburg, Malaysia, the Netherlands, Nigeria, Norway, Poland, Portugal, Rumania, Sierra Leone, South Korean, Spain, Sweden, Switzerland, Uganda, Turkey, United Kingdom, USA, New Zealand.

https://z-inspection.org

Our Mission

With Z-Inspection® we want to help to establish what we call a Mindful Use of AI (#MUAI).

Motivation of our work

How to asses Trustworthy AI in *practice*?

photo RVZ

Z-inspection® Process

We created a *participatory process* to help teams of skilled experts to assess the ***ethical, technical, domain specific*** **and** ***legal*** implications of the use of an AI-product/services within given *contexts*.

- **Published in IEEE Transactions on Technology and Society VOL. 2, NO. 2, JUNE 2021**

"Responsible use of AI" Pilot Project

- ***Pilot Project with the Province of Fryslan, Rijks ICT Gilde & the Z-Inspection® Initiative.***
- Marjolein Boonstra, Frédérick Bruneault, Subrata Chakraborty, Tjitske Faber, Alessio Gallucci, Eleanore Hickman, Gerard Kema, Heejin Kim, Jaap Kooiker, Elisabeth Hildt, Annegret Lamade, Emilie Wiinblad Mathez, Florian Möslein, Genien Pathuis, Giovanni Sartor, Marijke Steege, Alice Stocco, Willy Tadema, Jarno Tuimala, Isabel van Vledder, Dennis Vetter, Jana Vetter, Magnus Westerlund, Roberto V. Zicari.

- The pilot project took place from May 2022 through January 2023.
- During the pilot, the practical application of a deep learning algorithm from the province of Fryslân was assessed.

Environmental Monitoring

- The AI maps heathland grassland by means of satellite images for monitoring nature reserves. **Environmental monitoring** is one of the crucial activities carried on by society for several purposes ranging from maintaining standards on drinkable water to quantifying the CO2 emissions of a particular state or region.
- Using satellite imagery and machine learning to support decisions is becoming an important part of environmental monitoring.

Share the experiences, results and lessons learned from performing

- **A Trustworthy AI assessment** using the Z-Inspection® process and the EU framework for Trustworthy AI, combining it with
- **A Fundamental Rights assessment** using the Fundamental Rights and Algorithms Impact Assessment (FRAIA) as recommended by the Dutch government for the use of AI algorithms by the Dutch public authorities.

The "Assessment for Trustworthy AI" pilot sought to answer to the following questions

- As a government, how do you govern the development and use of *responsible* AI?
- What frameworks, laws and regulations are important, and how do we assess them in the development and use of AI?
- How do you analyze, assess and improve AI applications?
- And are the applications consistent with public values and human rights?
- What ethical issues does the AI system raise?
- What fundamental rights could be affected by the AI system?
- What measures could be met for the AI system to be *trustworthy*?

- The pilot gave some answers to these questions and in addition helped to stimulate awareness and dialogue about AI within the Dutch government, and provided guidelines to be able to confidently deploy AI technology for the questions of tomorrow.

- **"The results of this pilot are of great importance for the entire Dutch government, because we have developed a best practice with which administrators can really get started, and actually incorporate ethical values into the algorithms used."**
  – Rijks ICT Gilde - Ministry of the Interior and Kingdom Relations (BZK)

The background

- The Province of Fryslân is investing, in the coming years, in the smart and effective use of data. The province sees that almost all provincial developments and social tasks contain a data component. This creates urgency in the subject. To be able to responsibly respond to technological developments as a province, a sharp vision on data and AI is needed. Participation in the pilot helped design the future digital infrastructure and outline ethical frameworks.

- As directly related to this pilot, the Province of Fryslân is required by law to monitor biodiversity in natural areas. This is done by conducting a manual, visual inspection once every 10 years. There is a need to monitor and map the natural areas more often and monitor heather fields for grassification of heathlands and faster. To facilitate the process, reduce its costs and streamline the procedure, the Province commissioned a third party to develop an AI system for this purpose.

Scope of the Pilot

- The scope of the pilot was to assess whether the use of this AI system is trustworthy, which fundamental human rights are affected by the AI system, and how it can be used responsibly in practice.

Aim of the AI System

- The aim of the AI system is to **help ecologists** to quickly and frequently image the natural area so that it can be checked whether the intended nature quality objectives are being met, the right management measures can be taken and whether the approach to increasing biodiversity is working.
- Specifically the AI system **aims to provide information about the diffusion of the invasive and unwanted grass species** Molinia caerulea, known as moor grass or pipestraw, and Avenella flexuosa (common name wavy-hair grass) in heather fields using satellite images. The satellite images are made available by The Netherlands Space Office (NSO) on the free Satellite Data Portal where generic high resolution optical satellite images are available to be used in GIS.

Approach

In the pilot, the AI system was examined from three different perspectives:

- 1. Technical
- 2. Ecological
- 3. Ethical and Fundamental Human Rights

Ethics and Fundamental Rights Assessment

In light of the introduction of a fundamental rights impact assessment tool for algorithms in the Netherlands in March 2022, a hybrid approach was adopted in the pilot.

- First, the AI system was assessed against the **human rights** requirements using **the FRAIA**. This assessment did not only consider human rights violations but also rights which could be protected or strengthened by applying the AI system, such as the right to a healthy environment.
- Then **ethical issues** were identified and assessed based on the **European guidelines for Trustworthy AI** and the system was assessed from this broader perspective.

Claim, Arguments and Evidence

Our approach is unique, as it combines a fundamental rights assessment with a Trustworthy AI Assessment using a **evidence based approach**, using a framework called *Claim, Arguments and Evidence.*

Focus: fundamental rights

- In identifying the fundamental rights being affected by the AI system the assessment looked at the list of fundamental rights provided in the FRAIA, which are clusters around four groups with specific rights listed under each of the areas. Rights related to:
- The person
- freedom-related fundamental rights
- equality rights
- procedural fundamental rights

Focus: Ethics

- Following this step, the assessment considered more broadly ethical issues arising from the AI system. Specifically, the ethics guidelines for trustworthy artificial intelligence were considered as defined by the EU High-Level Expert Group on AI (AI HLEG, 2019). The four ethical principles of the guidance were used, acknowledging that tensions may arise between them:
- (1) Respect for human autonomy (2) Prevention of harm (3) Fairness (4) Explicability
- Furthermore, the seven requirements of Trustworthy AI defined in (AI HLEG, 2019) were considered. Each requirement has a number of sub-requirements .

Four Ethical Principles of the EU Trustworthy AI Framework

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy
(ii) Prevention of harm
(iii) Fairness
(iv) Explicability

- There may be **tensions** between these principles.

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

23

EU Seven Requirements and Sub-requirements for *Trustworthy* AI

24

source: *Ethics Guidelines for Trustworthy AI.* Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

- Requirements Sub-Requirements
- 1 Human agency and oversight *Including fundamental rights, human agency and human oversight*
- 2 Technical robustness and safety *Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility*
- 3 Privacy and data governance *Including respect for privacy, quality and integrity of data, and access to data*
- 4 Transparency *Including traceability, explainability and communication*
- 5 Diversity, non-discrimination and fairness *Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation*
- 6 Societal and environmental wellbeing *Including sustainability and environmental friendliness, social impact, society and democracy*
- 7 Accountability *Including auditability, minimization and reporting of negative impact, trade-offs and redress.*

Human Rights Assessment with a Evidence base approach

In this pilot, for each of the rights identified as potentially affected, the assessment concludes with a ***claim*** whether the right is

- a) affected (regardless of whether this is positively or negatively affected),
- b) not affected, or
- c) might be affected depending on certain clarifications.

A brief argument is made in respect of each claim and ***evidence*** is provided in support of whether the right is affected. The assessment identified five fundamental rights clusters which were potentially affected by the AI system.

I. Rights related to the Person

I. Rights related to the Person:

- Rights related to Healthy living Environment
- Rights related to Personal identity/personality rights/personal autonomy
- Rights related to Protection of data and informational privacy rights
- Rights related to Territorial privacy

II Procedural Rights

II Procedural Rights

- 5. Rights related to Right to good administration

Additional relevant aspects that arise from the Trustworthy AI Assessment

- In addition to the ethically relevant aspects discussed in the context of the fundamental rights-based assessment, additional ethical issues were identified for reflection.

Ethical issues

- Transparency and lack of transparency
- Receiving relevant information
- Human agency and oversight
- Technical robustness and safety
- Justice and fairness
- Cost reduction
- Diversity and Inclusion
- Responsibility and Accountability
- Due diligence in decision-making

Comparing the Trustworthy AI assessment process with the fundamental rights-based FRAIA assessment tool

- While the FRAIA tool was useful and clear, the question about how to frame the human rights assessment nevertheless arose and more specifically, how to consider the fundamental rights as part of an assessment of trustworthiness and ethical reflections on an AI system.
- Should we consider the rights as they are defined in law and interpreted through the courts only?
- Or should the rights be considered more broadly, as part of the assessment, linking the rights to ethical principles beyond their narrower legal definition?
- If only the legal definitions are used, an assessment of whether specific legislation applies would be required.

A two-tiered, integrated approach.

- **If the human rights assessment is defined too narrowly it risks being an assessment separate from the ethical assessment, or the assessment of trustworthiness.**
- **If it is too broad, the human rights standards risk being watered down.**

A two-tiered, integrated approach, looking both at legal requirements and the broader ethical questions, could be envisaged, depending on the organizational set up and use case.

Lessons Learned from the two assessment approaches

- The fundamental rights assessment and the ethics assessment based on the Trustworthy AI guidelines go hand in hand; both approaches provide critical insights with regard to the AI use case.

Similarities and Differences

Reflecting on AI from an ethics perspective clearly overlaps with a fundamental rights assessment.

- Both ethics and fundamental rights are about norms and fundamental values held in society.
- As ethics reflection and ethics guidelines influence law, scholars from both fields must work together when thinking about the shaping of technology and its societal implications.
- Even though there are great similarities, there are several considerable differences between the two approaches.

Differences

- **A fundamental rights-based approach is more closely linked to existing law and focuses on aspects that are legally relevant and thus enforceable.**
- **Compared to this, an ethics-based approach is much broader and also more open to reflection on potential implications that may not be worth considering from a legal perspective.**
- For example, from an ethics perspective, personal autonomy, freedom of decision-making, and fairness were found to be concepts of clear relevance in the context of the pilot project's AI tool, whereas, from a rights-based perspective, rights related to personal autonomy in a strictly legal sense were considered not infringed by the AI tool.

Different perspectives

- **While a fundamental rights-based assessment focuses on whether fundamental rights are negatively affected or infringed, from an ethics perspective, both positive and negative implications of AI technology are considered.**
- In this pilot, for example, the potential positive implications of the AI tool on the environment proved to be central.

Final Note

- Approaching the use case from a fundamental rights perspective implies that ethical and societal aspects and implications of AI are discussed only in-sofar as they are related to fundamental rights and existing law.

For more information

Report with the results
Lessons Learned in Performing a Trustworthy AI and Fundamental Rights Assessment.
Cite: arXiv:2404.14366 [cs.CY]
https://arxiv.org/abs/2404.14366

YouTube Video:
https://www.youtube.com/watch?v=z_RCysclXdk

Winner of the 2025 ISSIP Awards Excellence in Service Innovation with Distinguished Recognition – Impact to Society.
https://z-inspection.org/wp-content/uploads/2025/03/ISSIP2025-certificate.pdf

[발표 3 | Speaker 3]

신기술과 인권 대응: 대한민국 AI 법제와 인권 보장을 위한 과제

Responding to Emerging Technologies and Human Rights: Tasks for Safeguarding Human Rights under Korea's AI Legal Framework

안진현
AHN Jin-hyun

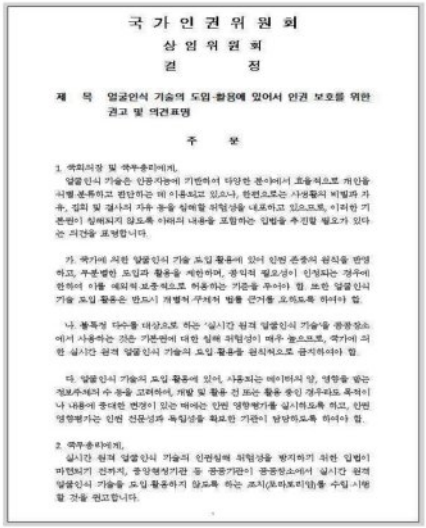
국가인권위원회 인권정책과 사무관
Deputy Director, Human Rights Policy Division, NHRCK

신기술과 인권 대응: 대한민국 AI 법제와 인권 보장을 위한 과제

안진현 | 국가인권위원회 인권정책과 사무관

목차

1. 신기술과 인권 : 국가인권위원회 관점에서 본 주요 쟁점



국 가 인 권 위 원 회
상 임 위 원 회
결 정

제 목 얼굴인식 기술의 도입·활용에 있어서 인권 보호를 위한 권고 및 의견표명

주 문

1. 국회의장 및 국무총리에게,
얼굴인식 기술은 인공지능에 기반하여 다양한 분야에서 효율적으로 개인을 식별 분류하고 판단하는 데 이용되고 있으나, 한편으로는 사생활의 비밀과 자유, 집회 및 결사의 자유 등을 침해할 위험성을 내포하고 있으므로, 이러한 기본권이 침해되지 않도록 아래의 내용을 포함하는 입법을 추진할 필요가 있다는 의견을 표명합니다.

가. 국가에 의한 얼굴인식 기술 도입·활용에 있어 인권 존중의 원칙을 반영하고, 무분별한 도입과 활용을 제한하며, 공익적 필요성이 인정되는 경우에 한하여 이를 예외적·보충적으로 허용하는 기준을 두어야 함. 또한 얼굴인식 기술 도입·활용은 반드시 개별적 구체적 법률 근거를 요하도록 하여야 함.

나. 불특정 다수를 대상으로 하는 '실시간 원격 얼굴인식 기술'을 공공장소에서 사용하는 것은 기본권에 대한 침해 위험성이 매우 높으므로, 국가에 의한 실시간 원격 얼굴인식 기술의 도입·활용을 원칙적으로 금지하여야 함.

다. 얼굴인식 기술의 도입·활용에 있어, 사용되는 데이터의 양, 영향을 받는 정보주체의 수 등을 고려하여, 개발 및 활용 전 또는 활용 중인 정주라도 목적이나 내용에 중대한 변경이 있는 때에는 인권 영향평가를 실시하도록 하고, 인권 영향평가는 인권 전문성과 독립성을 확보한 기관이 담당하도록 하여야 함.

2. 국무총리에게,
실시간 원격 얼굴인식 기술의 인권침해 위험성을 방지하기 위한 입법이 마련되기 전까지, 중앙행정기관 등 공공기관이 공공장소에서 실시간 원격 얼굴인식 기술을 도입·활용하지 않도록 하는 조치(모라토리엄)를 수립·시행할 것을 권고합니다.

개인정보 보호와 프라이버시

AI 발전은 대규모 데이터의 수집과 분석을 기반으로 하지만,
그 과정에서 개인정보와 사생활 침해 위험 증가

얼굴인식 AI, CCTV 등은 공공장소에서 불특정 다수의 생체정보를 무차별적으로 수집·감시할 수 있어 헌법상 사생활의 비밀과 자유를 침해할 가능성이 있음

대규모 감시 기술은 시민들로 하여금 자신의 행동이 기록·분석될 수 있다는 우려 속에 표현이나 집회 참여를 주저하게 만드는 '위축 효과'를 초래할 수 있음

1. 신기술과 인권 : 국가인권위원회 관점에서 본 주요 쟁점

"AI로 채용했다가 5억 날렸다"…소송 휘말리더니 '날벼락'
[김대영의 노무스쿨]

"AI로 직원 뽑는 건…" 불법 우려에 고개 젓는 기업들

AI 채용 둘러싼 분쟁 사례 '주목'
국내선 사례 없지만 대응책 필요
AI 채용 도입 기업 5곳 중 1곳꼴
채용절차법 등 현행법 위반 우려
AI 채용 도입 땐 법적 검토 필수

<출처 : 환경닷컴>

데이터 편향 등 차별

AI 시스템은 학습 데이터의 편향 등으로 인해
차별적인 결과를 초래할 수 있음

미국 강사를 고용하여 온라인 과외 서비스를 제공하는 중국 기업이
강사 채용 과정에서 활용한 인공지능 채용 서비스가 고령 지원자의 이력서를
자동으로 걸러 내어 200명 이상의 고령 지원자들이 채용기회를 박탈당함

미국 평등고용기회위원회(EEOC)가 2022. 5. 고용상 연령차별금지법 위반으로
이를 제소하였고, 2023. 8. 피해를 입은 지원자들에게 총 36만 5천달러
(약 5억 7백만원)을 배상, 피해자들에게 재지원 기회 허용 등 조건으로 합의

1. 신기술과 인권 : 국가인권위원회 관점에서 본 주요 쟁점

인공지능기본법

✓ 제3조 제2항
"영향받는 자는 인공지능의 최종결과 도출에 활용된
주요 기준 및 원리 등에 대하여 기술적 합리적으로 가능한
범위에서 명확하고 의미 있는 설명을 제공받을 수 있어야 한다"

✓ 제31조 제1항
"인공지능사업자는 고영향 인공지능이나 생성형 인공지능을
이용한 제품 또는 서비스를 제공하려는 경우 인공지능에
기반하여 운용된다는 사실을 이용자에게
사전에 고지하여야 한다"

투명성과 설명 가능성

AI 의 의사결정 과정이 불투명해 결과의 근거를 알기 어렵고,
책임 소재를 규명하기 힘든 문제가 있음

'블랙박스' AI는 사람이 이해하기 어려운 방식으로 판단을 내리므로,
권리 침해나 차별이 발생하더라도 당사자가 어떤 이유로 불이익을
받았는지 알지 못해 권리구제를 포기하는 상황이 발생

특히, AI 결정이 채용 탈락, 대출 거절, 수사 대상 선정 등
개인의 핵심 권리에 영향을 미칠 때 더욱 두드러짐

1. 신기술과 인권 : 국가인권위원회 관점에서 본 주요 쟁점

배달플랫폼라이더, AI 알고리즘 뒤 플랫폼업체 착취 밝히다

A 서비스연맹 | ⓒ 승인 2023.10.24 18:18

서비스연맹, 배달플랫폼노조_라이더 1030명 참여 실태조사 결과 발표
플랫폼사, AI 알고리즘과 약관 동의 강요로 노동환경 악화 강요
부당계약, 임금삭감, 지역차별 막을 제도 신설 절실해

코로나 팬데믹 동안 확장하던 배달플랫폼 산업이 다소 위축되면서 배달플랫폼라이더들의 노동환경이 더욱 열악해지고 있다. 대표적 배달플랫폼기업 배달의민족이 2년 연속 산재 승인 1위에 오르는 등 중대산업재해 지형까지 바꿔 놓고 있는 실정이다. 이에 배달플랫폼라이더 노동환경의 문제를 객관적, 심층적으로 분석하고 대책을 찾는 토론회가 개최됐다.

노동권과 일자리 영향

AI의 자동화 기술은 일자리 감소뿐 아니라 고용 불안정을 심화시키며 노동자의 권리를 위협할 수 있음

플랫폼 노동에서는 AI 알고리즘이 노동시간과 임금 등을 결정하면서 노동권 사각지대를 형성하고 있다는 우려 제기

AI 기반 채용, 인사평가, 해고 절차는 고용 안전성을 약화시키며, 노동자 감시 시스템은 사생활의 비밀과 자유를 침해할 위험도 지적됨

2. 주요 이해관계자들의 노력 : 국가인권위원회의 선도적 대응

<AI 개발과 활용에 관한 인권 가이드라인>

22년 5월 가이드라인 마련 후 정부에 활용 권고

- ✓ 인간의 존엄성과 존중
- ✓ 투명성과 설명 의무
- ✓ 자기결정권의 보장
- ✓ 차별금지
- ✓ 인공지능 인권영향평가 시행
- ✓ 위험도 등급 설정 및 관련 법과 제도 마련 등

<AI 인권영향평가 도구>

24년 5월 평가도구 마련 후 정부에 활용 의견표명

① AI 기술 관련 영향 분석 및 평가
- ► 개인정보보호, 데이터 관리
- ► 알고리즘의 신뢰성, 차별금지
- ► 설명가능성과 투명성
- ► 자동화 정도와 인간의 개입

② 인권에 미치는 영향 및 심각도
③ 위험에 대한 방지·완화 및 구제
④ 이해관계자의 참여
⑤ 평가 결과의 공개 및 평가에 대한 점검 등

2. 주요 이해관계자들의 노력 : 제도화와 공론화

2020년 7월	2023년 7월	2024년 12월	2025년 10월(예정)	2026년 1월
국회	국가인권위원회	국회	과학기술정보통신부	
인공지능 법률안 발의	인공지능 법률안에 대한 의견표명 → 제11조(우선허용 · 사후구제 원칙) 삭제 등 일부 의견 반영	인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 (인공지능기본법) 여야 합의로 통과	시행령, 가이드라인(고시) 마련 → 입법예고	인공지능 기본법 시행
시민 사회	✓ 정부 AI 정책·법안 지속 모니터링 ✓ 인권 관점에서 입법 결함 비판 및 개선 요구			
학계	✓ AI 윤리 · 법제 · 사회영향 · 기술 해결책 등 학제 간 연구 ✓ 이론적 틀과 실증 연구를 통해 정책 근거 제공			

3. 한국 AI기본법의 제정과 인권 가치 반영

1) EU / 한국 / 일본의 AI법 체계

✓ EU : 세계 최초 포괄적 AI법으로 AI 위험을 기준으로 등급을 구분하여 차등 규제 적용, 법적 구속력 있는 규제체계
✓ 한국 : 세계 두번째 포괄적 AI법으로 등급 구분 없이 고영향 AI에 대해 규제, EU와 일본의 AI법의 중간적 성격
✓ 일본 : AI 연구개발, 이용의 촉진을 뒷받침하기 위한 진흥법 마련, 연성 규제체계로 정부는 조사와 지도 권한을 갖고 기업은 협력 의무 부담

구분	EU(AI ACT)	한국(AI기본법)	일본(AI추진법)
법의 목적	안전 + 기본권 보호	권익과 존엄성 보호 + 국가경쟁력 강화	국민생활 + 경제발전
시행 시기	24년 1월 제정 → 25년 금지 AI 및 범용 AI 시행 → 26년 8월 전체 시행	24년 12월 국회 통과 → 26년 1월 시행	25년 6월 시행
규제방식	금지 / 고위험 / 제한적 / 저위험 AI로 구분하여 '금지 / 고위험'에 대한 규제 명시 * 제한적 AI 투명성 의무 부과	등급 구분 없이 고영향 AI에 대한 규제 명시 생성형 AI 투명성 의무 부과	없음
사업자 의무	위험도별 차등 의무 부과 행위자별 차등 의무 부과	투명성·안전성 확보 의무 고영향 AI 확인·사업자 책무 고영향 AI 기본권 영향평가 노력 의무	협력 의무
제재 수준	고위험 AI 과징금 전년 매출 3%, 1천 5백만 유로 중 높은 금액	3천만원 과태료	없음

3. 한국 AI기본법의 제정과 인권 가치 반영

2) 인권 관점에서의 강점과 한계

인권 관점에서의 강점		인권관점에서의 한계
법의 목적에 '국민의 권익과 존엄성 보호' 명시		인권영향평가 실효성 부족 - 고위험 AI 기본권 평가 '노력 의무' 수준
AI 신뢰성 확보를 위한 사업자의 책임과 의무 부여	VS	AI로 인한 피해구제 권리보장 미흡
책임성 강화 노력 -AI기본법 위반사항 조사 · 시정명령		산업 진흥 중심 구조 감독주체가 산업부처에 집중 → 이해 상충 우려
파편적 규제 → 구조화된 거버넌스로 전환		명확한 금지조항 부재 - EU AI ACT의 허용 불가 위험 규정과 대비

4. 인권 가치 반영을 위한 AI 법 · 정책 방향

1) 공공부문 AI 인권영향평가 의무화

UN · 국제사회	캐나다, 유럽 등에서 정부 기관이 AI를 도입 · 활용할 때 AI 영향평가를 수행하도록 하는 추세
인권위 의견표명(23.7.13.)	AI 개발 · 출시 전 인권영향평가 실시 → 수정 및 활용 범위 변경 시 재평가
AI기본법	고영향 AI 기본권 영향평가가 '노력 의무'에 그쳐 실효성이 저해될 우려

▼

인권 가치 반영

- 공공영역에서 AI가 국민의 권리와 의무에 미치는 영향이 크다는 특수성을 고려하여 우선적으로 공공기관 AI 인권영향평가 의무화 필요
- 인권영향평가 → 단순한 윤리기준을 넘어서 인공지능에 대한 인권기반 접근 필요

25년 8월 25일, 데이터기반행정법 일부개정법률안 국회 발의

✓ 인공지능 활용에 따른 최종 권한과 책임이 해당 공공기관에 있음을 명확히 하도록 책무를 추가(안 제3조제5항)

✓ 공공기관의 장은 인공지능 도입 전에 인공지능 영향평가를 실시하고 그 결과를 공표해야(안 제32조)

4. 인권 가치 반영을 위한 AI 법 · 정책 방향

2) AI로 인한 피해구제 절차 마련

인권위 의견표명(23.7.13.)	AI의 영향을 받는 자의 권리를 명시하고, 피해 발생 시 구제 절차 마련
AI기본법	설명 제공 등 영향받는 자의 권리 일부 명시, 권리구제 조항은 없음

▼

인권 가치 반영

- 사회전반에 AI 활용 본격화, AI 오작동, 데이터 편향 등으로 인권침해나 차별 발생 가능성 증가 우려
- 권리 침해 유형별 분류 → 기존 구제 제도에 근거 규정 마련 필요
- 새로운 유형의 권리 침해 대응을 위한 별도 구제 절차 마련 필요

24년 3월 15일, 개인정보보호법 자동화된 결정에 대한 정보주체의 권리 조항 시행

✓ 자신의 권리 또는 의무에 중대한 영향을 미치는 경우, 자동화된 결정을 거부하거나 결정에 대한 설명 요구 등(제37조의2)

4. 인권 가치 반영을 위한 AI 법 · 정책 방향

3) AI 감독 및 규제, 산업 진흥 부처와 분리

인권위 의견표명(23.7.13.)	AI 감독 및 규제에 관한 사항을 독립적인 기관이 담당 △인공지능 위험 요소 점검 및 감독, △인공지능 인권영향평가, △인공지능으로 인한 피해구제
AI기본법	AI 위험 요소 점검 및 감독 산업 진흥 부처(과기정통부)가 담당

▼

인권 가치 반영

- 산업 진흥과 규제 업무를 하나의 기관에서 모두 담당할 경우, 규제의 실효성 저해 우려
- AI 감독 · 규제 업무는 AI 산업 진흥에 관한 사항을 소관하는 기관이 아닌 제3의 기관이 독립적으로 수행 필요

유엔인권이사회 '디지털 시대 프라이버시권 보고서'

✓ AI 시스템에 대한 적정하고 독립적이고 공정한 감독이 필요하고,

✓ 이는 행정적 사법적 준사법적 기관 및 의회 감독 기관의 조합으로 수행될 수 있으며,

✓ 개인정보 보호 기관, 소비자 보호 기관, 부분별 규제 기관, 차별 방지 기구 및 국가인권기구가 감독 시스템의 일부로서 구성되어야 한다.

4. 인권 가치 반영을 위한 AI 법 · 정책 방향

4) 위험성 수준과 AI 사업자 유형에 따른 규제 세분화

인권위 의견표명(23.7.13.)	AI을 적정한 등급으로 구분하고, 각 등급에 맞게 규제 수준으로 다르게 규정
AI기본법	등급 구분 없이 고영향 AI 에 대해 규제

▼

인권 가치 반영

- 인공지능 활용 분야 · 영역의 특성, 해당 분야 · 영역에서 AI가 국민의 인권과 안전에 미치는 영향 및 위험성 고려하여, 위험도별 차등 규제 방안 검토
- AI개발사업자와 AI이용사업자의 책무를 구분하여 법령에 명시하는 방안 검토

EU AI ACT

✓ 금지 / 고위험 / 제한적 / 저위험 AI로 구분하여 주로 '금지 / 고위험' 대상으로 엄격한 의무 부과

✓ AI제공자(provider)가 해당 AI에 관하여 가장 많은 정보를 보유하게 되므로 전반적으로 '공급자' 중심으로 의무 부과

✓ AI배포자(deployer)에게 이용 관련 일정 의무 부과

Responding to Emerging Technologies and Human Rights: Tasks for Safeguarding Human Rights under Korea's AI Legal Framework

AHN Jin-hyun | Deputy Director, Human Rights Policy Division, NHRCK

1. New Technologies and Human Rights: Key Issues from the Perspective of the National Human Rights Commission of Korea (NHRCK)

(1) Personal Data Protection and Privacy

The development of artificial intelligence (AI) is grounded in the large-scale collection and analysis of data, which inevitably heightens the risks of personal data breaches and intrusions into private life. Technologies such as facial recognition AI and surveillance cameras can indiscriminately collect and monitor the biometric information of unspecified individuals in public spaces, thereby posing a potential threat to the constitutional right to privacy and the right to private life.

AI models trained on sensitive data may profile individuals and, without their consent, infer information such as political beliefs, health conditions, or sexual orientation. This poses significant risks of violating the right to privacy. Furthermore, large-scale surveillance technologies can generate a chilling effect, whereby citizens refrain from exercising their freedom of expression or participating in assemblies out of concern that their behavior may be recorded and analyzed.

(2) Algorithmic Bias and Discrimination

AI systems can produce discriminatory outcomes due to biases embedded in training data or limitations in the development process. Indeed, there have been reported cases abroad where AI systems, used in recruitment processes, disadvantaged applicants based on factors such as age or race, which subsequently led to litigation.

(3) Lack of Transparency and Explainability

AI decision-making processes often lack transparency, making it difficult to identify the basis of outcomes or to assign accountability. So-called "black-box" AI systems operate in ways that are not readily comprehensible to humans. As a result, even when rights violations or discriminatory outcomes occur, affected individuals may not know why they were disadvantaged and may ultimately forgo seeking remedies. This poses a serious challenge to procedural justice and accountability, particularly when AI decisions affect fundamental rights—such as rejection in recruitment, denial of loans, or being targeted for investigation. Ensuring explainability has therefore emerged as a new pillar of human rights protection in the age of AI.

(4) Labor Rights and the Impact on Employment

AI and automation technologies not only reduce the number of jobs but also exacerbate employment insecurity, thereby threatening workers' rights. In particular, concerns have been raised in the context of platform labor, where AI algorithms determine working hours, wages, and other conditions, creating gaps in the protection of labor rights. Moreover, AI-based recruitment, performance evaluations, and dismissal procedures can undermine job security, while worker surveillance systems pose risks to the constitutional right to privacy and private life.

2. Efforts by Key Stakeholders: Institutionalization and Public Discourse

(1) The NHRCK's Pioneering Response

The National Human Rights Commission of Korea (NHRCK) has taken a leading role in shaping Korea's discourse on AI and human rights. In May 2022, it introduced the Human Rights Guidelines on the Development and Use of Artificial Intelligence, which set out human rights principles to be applied throughout the development and deployment of AI. The NHRCK further recommended that the government use these Guidelines as the foundation for AI-related policies and legislation. The Guidelines encompass principles such as respect for human dignity, transparency and the duty of explanation, protection of the right to self-determination, non-discrimination, the implementation of human rights impact assessments on AI (AI HRIA), the establishment of risk-based classifications, and the introduction of related legal and institutional frameworks.

In May 2024, the NHRCK also developed the Human Rights Impact Assessment Tool for Artificial Intelligence to help prevent human rights violations and discrimination caused by AI, and expressed its opinion that the government should adopt this tool. The tool is composed of 72 assessment items across four stages, grounded in the Constitution of the Republic of Korea, UN international human rights treaties, and guidance from the UN Human Rights Council.

(2) The National Assembly's efforts for legislation

To strike a balance between fostering the AI industry and establishing a foundation for trustworthiness, the National Assembly engaged in nearly four years of deliberations and, in December 2024, passed the Framework Act on the Development of Artificial Intelligence and the Establishment of Foundation for Trustworthiness (commonly referred to as the Framework Act on Artificial Intelligence) with bipartisan support. The Act was promulgated on January 21, 2025, and will enter into force on January 22, 2026. With this, Korea became the second jurisdiction in the world, after the European Union, to establish a comprehensive legal framework on AI.

During the legislative review process in July 2023, the NHRCK expressed its opinion on the draft bill, emphasizing the need for provisions to prevent human rights violations and discrimination. Some of these recommendations were incorporated into the final version of the Act.

(3) Government Policies and Efforts for Public Discourse

Since January 2025, the Ministry of Science and ICT (MSIT) has been preparing subordinate legislation—including enforcement decrees, public notices, and guidelines—to facilitate the timely implementation of the Framework Act on the Development of Artificial Intelligence and the Establishment of Foundation for Trustworthiness. In particular, to support businesses in fulfilling their obligations regarding the minimum requirements for ensuring AI safety and trustworthiness, the MSIT is developing detailed guidelines (public notices) corresponding to Articles 31 through 35 of the Act.

Furthermore, with the establishment of the AI Safety Institute in November 2024, the government has strengthened risk management for AI and expanded international cooperation with countries such as the United States, the United Kingdom, and Japan. At the same time, it

has promoted public engagement to build a social foundation for AI safety and trustworthiness.

(4) Civil Society's Monitoring and Advocacy

Civil society organizations closely monitor the government's AI policies and legislative initiatives, critically pointing out legislative gaps from a human rights perspective. When they perceive the government as being overly focused on technological promotion, they serve as an oversight mechanism by demanding stronger human rights safeguards to protect the public's safety and rights from the risks posed by high-risk AI systems.

(5) Academic Research and Policy Recommendations

Academia and research institutions play a vital role by conducting interdisciplinary studies that encompass AI ethics, legal frameworks, social impacts, and technological solutions. Their contributions include developing theoretical frameworks, producing empirical research, proposing ethical guidelines, and offering policy recommendations.

3. Enactment of Korea's AI Framework Act and the Incorporation of Human Rights Values: Positive Aspects and Limitations

(1) Overview of the Act

Korea's AI Framework Act occupies a middle ground between the European Union's AI Act—which adopts strict regulations with the aim of ensuring safety and protecting fundamental rights—and Japan's AI Promotion Act, which emphasizes self-regulation to enhance industrial competitiveness. The AI Framework Act pursues dual objectives: the protection of the rights and dignity of the people, and improving people's well-being and strengthening Korea's global competitiveness. It seeks to promote technological innovation while imposing only the minimum necessary regulations to ensure trustworthiness and safety. However, opinions are divided as to whether the Act adequately incorporates human rights values.

(2) Positive Aspects from a Human Rights Perspective

The AI Framework Act explicitly states the "protection of the rights and dignity of the people"

in its purpose clause and imposes obligations on providers, including requirements to ensure transparency as well as provisions for the safety and trustworthiness of high-impact AI. It also authorizes the MSIT to investigate violations and issue corrective orders. These measures are regarded as a first step toward responsible AI governance. Moreover, the very fact that a comprehensive legal framework on AI has been established carries significance in that it moves beyond fragmented regulations toward a structured form of governance, while also recognizing the broader social impacts of AI.

(3) Limitations from a Human Rights Perspective

From a human rights perspective, many essential elements remain issues to be addressed. As it currently stands, the Act has been criticized for lacking sufficient effectiveness to prevent and control key risks, such as discrimination caused by AI, threats arising from surveillance, and safety-related hazards.

First, contrary to the NHRCK's recommendation, the requirement to conduct human rights impact assessments for high-risk AI remains only a best-efforts obligation, which significantly weakens its preventive function.

Second, the Act contains no provisions that explicitly prohibit AI practices that should be banned outright. Internationally, instruments such as the EU AI Act classify AI systems that undermine human dignity—for example, AI designed to manipulate behavior or decision-making, exploit socially vulnerable groups, or indiscriminately harvest facial images—as posing "unacceptable risks" and prohibit their use. By contrast, Korea's AI Framework Act does not include such prohibitions.

Third, there are gaps in terms of remedies and the protection of rights. Although the Act defines "persons affected by AI," it does not specify provisions on their rights or the remedies available to them.

Fourth, the Act has been criticized for placing greater emphasis on industrial promotion than on human rights protection, thereby limiting its effectiveness. In particular, supervisory authority is concentrated in the MSIT, the ministry responsible for promoting industry, raising concerns about potential conflicts of interest.

4. The Way Forward for AI Laws and Policies to Incorporate Human Rights Values

(1) Mandating Human Rights Impact Assessments for AI in the Public Sector

At present, human rights impact assessments remain only a best-efforts obligation and thus lack effectiveness. The UN and many countries are increasingly paying attention to the negative impacts of AI in the public sector, as well as high-risk AI in the private sector, and are moving beyond simple ethical standards toward various forms of impact assessments aimed at preventing and managing such risks in advance. Given the opacity and autonomy inherent in AI systems, ex-post remedies are difficult to achieve, making it necessary to introduce preventive human rights impact assessments as a legal obligation.

(2) Differentiating Regulation by Risk Level and Type of AI Provider

It is necessary to refine laws and institutions so that the degree of risk in AI—classified into categories such as prohibited, high-risk, limited-risk, and low-risk—can be matched with appropriate regulatory measures and levels of human oversight, taking into account both the risk level and the type of AI provider. Such an approach is expected to protect public safety and fundamental rights while not hindering innovation, and to help build public trust in AI, thereby contributing to the creation of a sustainable AI ecosystem.

(3) Establishing Remedies for Affected Individuals

With the advancement of AI technologies, the likelihood of unforeseen harms to individuals has increased—such as those caused by system malfunctions, discrimination resulting from biased training data, or the unpredictability of algorithmic outcomes.

As things stand, there are no clear provisions ensuring that individuals harmed by AI can effectively obtain remedies. To address this, it is necessary not only to analyze and categorize different types of rights violations, but also to identify which existing remedy mechanisms are most appropriate for addressing each category. Legal grounds should then be established to enable the effective application of those mechanisms, accompanied by revisions to relevant laws. In addition, for types of rights violations that cannot be adequately addressed through existing procedures, there is a need to establish separate remedy mechanisms that take into account the unique characteristics of AI.

(4) Establishing an Independent Supervisory Body

To effectively prevent and oversee issues such as human rights violations and discrimination caused by AI, it would be desirable for an independent third-party —rather than a ministry tasked with promoting industry—to take exclusive responsibility.

세션 2
Session 2

신기술과 인권 과제: 불평등과 차별의 문제
Challenges of New Technologies: Inequality, Exclusion, and Human Rights

사회자 Moderator

김민호 | 성균관대학교 법학전문대학원 교수
KIM Minho | Professor, School of Law, Sungkyunkwan University

발표자 Speaker

이권일 | 경북대학교 법학전문대학원 부교수
LEE Kwon il | Associate Professor, School of Law, Kyungpook National University

팀 엥겔하르트 | 유엔 인권최고대표사무소 인권담당관
Tim ENGELHARDT | Human Rights Officer, OHCHR

나이갓 다드 | 디지털 권리 재단 상임이사, 유엔 AI 자문기구 위원
Nighat DAD | Executive Director, Digital Rights Foundation Member, UN High-level Advisory Body on AI

마이클 키웻 | 요하네스버그대학교 사회변화연구센터 선임연구원, 예일대학교 방문 교수
Michael KWET | Senior Researcher, University of Johannesburg Visiting Professor, Yale University

이승윤 | 중앙대학교 사회복지학과 교수
LEE Seung yoon | Professor, Department of Social Welfare, Chung-Ang University

세션 2 | Session 2

[사회자_Moderator]

김민호

KIM Minho

성균관대학교 법학전문대학원 교수

Professor, School of Law, Sungkyunkwan University

[주요경력]

1998년부터 성균관대학교 법학전문대학원 교수로 재직 중이다. 성균관대학교에서 학사, 석사, 박사학위를 받고, 미국 Boston University Law School에서 박사후연구과정(Post Doc.)을 수료했다.

개인정보보호법학회 회장, 국가인권위원회 인권위원, 대통령소속 규제개혁위원회 위원, 중앙행정심판위원회 위원 등을 역임했다.

지금은 공공데이터분쟁조정위원회 위원장, 국가기준데이터위원회 위원장, 한국인터넷자율정책기구(KISO) 이사회 의장으로 활동하고 있다.

[Career]

Professor Kim Minho has been a professor at Sungkyunkwan University Law School since 1998. He holds a bachelor's, master's, and doctoral degree from Sungkyunkwan University, and completed a postdoctoral program at Boston University Law School.

He has served as the president of the Korea Institute of Information Security and Cryptology (KIISC), a commissioner of the National Human Rights Commission of Korea, a member of the Presidential Regulatory Reform Committee, and a member of the Central Administrative Appeals Commission.

He is currently the Chairperson of the Open Data Mediation Committee, the Chairperson of the National Standard Data Committee, and the Chairperson of the board of directors of the Korea Internet Self-governance Organization (KISO).

세션 2 | Session 2

[발표자_Speaker]

이권일

LEE Kwon il

경북대학교 법학전문대학원 부교수

Associate Professor, School of Law, Kyungpook National University

[주요경력]

이권일은 경북대학교에서 법학을 전공하고 독일 튀빙겐 대학교에서 공법학으로 박사(Dr.iur)학위를 취득하였다. 헌법재판소 헌법연구원으로 근무하였고 동아대학교 법학전문대학원에서 교수로 재직하였으며, 현재 경북대학교 법학전문대학원에서 헌법 과목을 강의하고 있다. 현대 인터넷 사회에서의 프라이버시 보호, 특히 개인정보보호에 관심이 많으며 최근에는 인공지능기술 발달에 대한 규범학의 대응에 대해 연구하고 있다.

[Career]

Kwon-il Lee majored in law at Kyungpook National University and obtained his doctorate (Dr.iur) in public law from the University of Tübingen in Germany. He worked as a constitutional researcher at the Constitutional Court of Korea and served as a professor at Dong-A University Law School. He is currently teaching constitutional law at Kyungpook National University Law School. He has a strong interest in privacy protection in modern internet society, particularly personal information protection, and is currently researching how legal studies should respond to the development of artificial intelligence technology.

세션 2 | Session 2

[발표자_Speaker]

팀 엥겔하르트

Tim ENGELHARDT

유엔 인권최고대표사무소 인권담당관

Human Rights Officer, OHCHR

[주요경력]

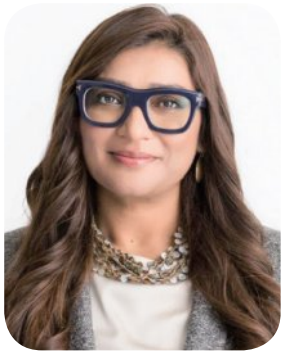
팀 엥겔하르트는 유엔 인권최고대표사무소(OHCHR)에서 신기술과 신흥 기술이 인권 향유에 미치는 영향에 대해 중점적으로 다루고 있습니다. 그는 디지털 감시, 온라인상에서의 자유로운 표현, 인공지능, 그리고 사이버 범죄를 포함한 다양한 관련 이슈들을 다루고 있습니다.
OHCHR에 합류하기 전에는 세계지식재산권기구에서 근무했으며, ICT 관련 사안을 전문으로 다루는 민간 법률 사무소에서도 일했습니다. 그는 베를린의 훔볼트 인터넷 법률 클리닉의 공동 설립자이기도 합니다.
그는 독일과 미국 변호사 자격을 모두 갖추고 있으며, 컬럼비아 로스쿨에서 법학 석사(LL.M.) 학위를, 취리히 대학교에서 박사 학위를 받았습니다.

[Career]

Tim Engelhardt's work at the UN Human Rights Office (OHCHR) focusses on the impacts of new and emerging technologies on the enjoyment of human rights. He covers a broad range of related issues, including digital surveillance, free expression online, artificial intelligence and cybercrime. Prior to joining OHCHR, he worked at the World Intellectual Property Organisation and in private legal practice, where he specialised in ICT related matters. He was a co-founder of the Humboldt Internet Law Clinic in Berlin. He has trained as a German and US lawyer and holds an LL.M. from Columbia Law School and a Ph.D. from Zurich University.

세션 2 | Session 2

[발표자_Speaker]



나이갓 다드

Nighat DAD

디지털 권리 재단 상임이사, 유엔 AI 자문기구 위원
Executive Director, Digital Rights Foundation
Member, UN High-level Advisory Body on AI

[주요경력]

나이갓 다드는 파키스탄 출신의 변호사이자 디지털 정책 전문가로, Digital Rights Foundation(DRF)을 설립해 이끌고 있습니다. 그녀는 유엔 사무총장 산하 인공지능 고위급 자문기구, 메타(페이스북) 감독위원회, 세계경제포럼 데이터 프런티어 글로벌 미래위원회 등 여러 국제 위원회에서 활발히 활동하고 있습니다. 그녀의 활동은 유엔 글로벌 디지털 콤팩트 등 국제 기술 거버넌스와 정책·규제 논의 전반을 포괄하며, 마이크로소프트와 메타 같은 글로벌 기업은 물론 영국, EU, 미국 정부에도 자문을 제공해 왔습니다.
또한 미국 전임 행정부 시절에는 백악관의 요청으로 전 세계적으로 확산되는 기술 기반 성폭력 문제 해결 방안에 대해 조언하기도 했습니다. 온라인 표현의 자유, 여성 권리, 플랫폼 책임성 분야에서 보여준 리더십은 국제적으로도 널리 인정받아, 타임(Time)지는 그녀를 "차세대 리더(Next Generation Leader)" 중 한 명으로 선정했습니다.

[Career]

Nighat Dad is a Pakistani lawyer, digital policy expert, and the founder of the Digital Rights Foundation (DRF). She serves on the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence, the Meta (Facebook) Oversight Board, and the World Economic Forum's Global Future Council on Data Frontiers, among several other international boards. Her contributions span international tech governance, policy, and regulation processes, including the United Nations Global Digital Compact, as well as providing expert recommendations and advice to companies such as Microsoft and Meta, and to governments including the UK, EU and US. In the previous U.S. administration, she advised the White House on addressing technologyfacilitated gender-based violence worldwide. Widely recognized for her leadership on online freedom of expression, women's rights, and platform accountability, she was named by TIME magazine as one of its Next Generation Leaders.

세션 2 | Session 2

[발표자_Speaker]

마이클 키웻

Michael KWET

요하네스버그대학교 사회변화연구센터 선임연구원, 예일대학교 방문 교수
Senior Researcher, University of Johannesburg
Visiting Professor, Yale University

[주요경력]

마이클 키웻 박사는 요하네스버그대학교 선임연구원이며 『디지털 디그로스: 생존 시대의 기술』과 『케임브리지 인종과 감시 핸드북』의 저자이다. 키웻 박사의 연구는 디지털 식민주의, 기술과 환경, 교육기술, 감옥 기술, 기술법 주제에 중점을 둔다. 그는 『뉴욕타임스』, 『알자지라』, 『인터셉트』, 『바이스 뉴스』, 『메일 앤 가디언』, 『트루스딕』에 기고했다. 퀘트 박사는 예일대학교 법학대학원에서 8년간 방문연구원으로 활동했으며, PeoplesTech.org 웹사이트의 창립자이다.

[Career]

Dr Michael Kwet is a Senior Researcher at the University of Johannesburg and author of the book, Digital Degrowth: Technology in the Age of Survival and The Cambridge Handbook of Race and Surveillance. Dr Kwet's research focuses on the topics of digital colonialism, tech and the environment, EdTech, carceral technologies, and tech law. He has published at The New York Times, Al Jazeera, The Intercept, VICE News, Mail & Guardian, and Truthdig. Dr Kwet was a Visiting Fellow at Yale Law School for 8 years, and is the founder of the website, PeoplesTech.org.

세션 2 | Session 2

[발표자_Speaker]

이승윤

LEE Seung yoon

중앙대학교 사회복지학과 교수

Professor, Department of Social Welfare, Chung-Ang University

[주요경력]

이승윤 교수는 중앙대학교 사회복지학부 교수로 영국 옥스퍼드대학교에서 사회정책학 박사학위를 취득했다. 교토대학교 사회학과와 이화여자대학교 사회복지학과에서 각각 조교수와 부교수로 재직한 경력이 있으며, 복지국가와 노동시장, 불안정 고용을 주요 연구분야로 하고 있다. 약 80편의 피어리뷰 학술논문과 저서를 발표했으며, 특히 2023년 출간한『Varieties of Precarity: Melting Labour and the Failure to Protect Workers in the Korean Welfare State(Policy Press) ((국문명)불안정의 다양성: 액화노동과 한국 복지국가의 노동자 보호 실패)』로 한국비판사회복지학회 대안연구상을 수상했다. 2020년부터 2022년까지 국무조정실 청년정책조정위원회 초대 민간 부위원장을 역임했고, 국민연금제도개혁위원회 위원으로 활동했다. 현재 '아시아 디지털화, 불안정노동, 복지국가 개혁' 연구프로젝트를 이끌고 있다.

[Career]

Professor Seung-yoon Lee is a professor in the Department of Social Welfare at Chung-Ang University, having earned her Ph.D. in Social Policy from the University of Oxford, UK. She previously served as an assistant professor in the Department of Sociology at Kyoto University and as an associate professor in the Department of Social Welfare at Ewha Womans University. Her main research areas include welfare states, labor markets, and precarious employment. She has published approximately 80 peer-reviewed academic papers and books, and notably received the Alternative Research Award from the Korean Critical Social Welfare Association in 2023 for her book "Varieties of Precarity: Melting Labour and the Failure to Protect Workers in the Korean Welfare State" (Policy Press). From 2020 to 2022, She served as the inaugural civilian vice-chairperson of the Youth Policy Coordination Committee under the Office for Government Policy Coordination, and was a member of the National Pension Reform Committee. She currently leads the research project on "Asian Digitalization, Precarious Labor, and Welfare State Reform."

[발표 1 | Speaker 1]

AI와 불평등의 재생산

AI and the Reproduction of Inequality

이권일

LEE Kwon il

경북대학교 법학전문대학원 부교수

Associate Professor, School of Law, Kyungpook National University

AI와 불평등의 재생산

이권일 | 경북대학교 법학전문대학원 부교수

I. 들어가며

4차산업혁명, 지능정보사회, 빅데이터, 인공지능(AI), 스마트시티, 스마트팜 등의 용어는 이제 더 이상 생경한 용어가 아니라 일반 국민에게도 너무나도 익숙해진 단어이다. IT 기술의 발달에 따른 사회의 변화는 더욱 가속화되고 있으며 그에 따라 우리의 생활도 변화하고 있다. 이 중 최근 가장 많이 대중화되고 논의되는 것이 인공지능 기술의 발전이다.

2016년 알파고와 이세돌 9단과의 대결은 우리에게 인공지능 기술의 발전에 관한 충격을 주었고, 2022년 말 생성형 AI 서비스의 대표적인 범용 서비스라고 할 수 있는 ChatGPT(Generative Pre-trained Transformer) 등의 서비스가 실제 시장에서 제공되고 그 기능이 급속도로 발전하는 모습을 경험함으로써 AI 기술의 상용화와 이로 인한 여러 발전 가능성과 문제점에 대한 논의가 뜨겁다. 특히 최근에는 AI가 상업적으로 많이 활용되고 인공지능 기술이 접목된 기기가 시장에서 반응이 좋기 때문에 구글, MS 등과 같은 글로벌 IT 사업자, 삼성과 애플 등의 회사는 물론 세계 각국에서 인공지능 기술 개발에 많은 인력과 비용을 투자하고 경쟁에서 앞서기 위해 사활을 걸고 있다. 이러한 인공지능의 범용화는 인공지능기술이 더 이상 과학자나 개발자, 또는 관련 전문가들끼리 논의되거나 베타 서비스로 개발되어 제공되던 단계를 넘어서서 우리의 삶에 깊숙이 관련된다는 점을 의미하는 중요한 변환점이라고 할 수 있겠다.

인공지능 기술의 발전은 우리에게 많은 편리함을 줄 것이라는 믿음과 함께 그 기술 발달의 결과를 예측할 수 없다는 우려와 공포를 주기도 한다. 또한 인공지능은 그 기술 발현의 과정(알고리즘)을 알기 어렵거나 알 수 없다(또는 공개할 수 없다)는 불투명성으로 인하여 여러 문제점이 제기되기도 한다. 이 글은 인공지능과 관련한 여러 문제점 중 인공지능의 편향성(공정성) 문제와 인공지능으로 인한 불평등의 재생산 문제를 다루고자 한다. 이를 위해 인공지능의 편향성으로 인한 차별의 예(II.)와 이러한 현상이 발생하는 이유(III.)에 대해 알아보고, 이를 개선하기 위한 방안(IV.)을 살펴보기로 한다.

II. 인공지능의 편향성으로 인한 차별의 예

인공지능이 인간과 유사한 사고체계, 지적능력을 가지고 있다고 하더라도 인간이 아닌 기계이다. 그리

고 우리는 지금까지 기계는 가치중립적인 것이란 믿음을 가지고 있었고, 따라서 인공지능이 기계인 이상 인간보다 공정할 것, 편향적이지 않을 것이라는 믿음이 있었다고 보여진다. 예를 들어 인간 개입없는 AI를 통한 네이버 뉴스편집[1], 인공지능 면접의 도입 등이 그러하다.

하지만 인공지능이 개발되어서 사용된 결과를 보았을 때 실제로는 인공지능의 편향성으로 인한 평등, 차별, 혐오 문제가 인공지능의 부작용 중 많은 부분을 차지함을 경험하고 있다. 주로 외국의 예를 우리가 보기 때문에 외국의 경우 우리와 차별문제의 양상이 다르다고 할 수도 있겠지만- 외국에서 주로 문제가 되는 것이 인종, 종교로 인한 차별 문제가 크게 부각되기 때문에- 우리의 경우도 성별에 의한 차별이나 성적 지향에 대한 차별, 나이, 학력, 지역에 의한 차별, 그리고 정치적 양극화로 인한 공정성 문제가 심각하여 이러한 문제에서 자유롭기 어렵다.

먼저 차별 양상을 분석하여 이에 대한 해결책을 모색하기 위해 인공지능으로 인한 차별의 대표적인 예와 양상을 살펴본다.

1. 성별과 관련한 차별

먼저 성별(젠더)과 관련된 차별을 살펴본다.

젠더와 관련하여 문제되는 것으로 AI의 성별 또는 성역할의 문제가 있다. 예를 들어 인공지능 비서의 경우 비서의 성별(이름과 음성 등)을 여성으로 하는 경우가 많다. SKT '누구', KT '기가지니', 네이버 '프렌즈', 카카오 '미니', 아마존 '알렉사', 마이크로소프트의 '코타나', 애플 '시리' 등의 인공지능 비서서비스는 여성의 목소리를 기본으로 설정함은 물론 이용자가 이를 변경할 수도 없는 경우도 많았다. 이러한 현상은 우리 사회의 성역할의 문제 또는 젠더의 문제를 그대로 반영함은 물론 젠더 편향을 강화할 수 있다는 우려를 일으킨다.[2]

또한 구글의 자동번역 서비스에서는 엔지니어, 의사와 같은 직업군은 남성, 간호사 등의 직업을 여성으로 번역하는 등의 기존 성역할 관행을 그대로 나타내기도 하였다.[3]

---

1 네이버 모바일 첫화면에서 뉴스 포기한다, 미디어오늘, 2018.05.09., https://www.mediatoday.co.kr/news/articleView.html?idxno=142622, 물론 이에 대해서도 공정성에 대한 문제점이 제기되었었다.

2 한애라, 인공지능과 젠더차별, 이화젠더법학 제11권 제3호, 2019; 허유선, 인공지능의 젠더 차별 사례 분석, 윤리교육연구 제71집, 2024, 461쪽 이하

3 이윤아, 윤상오, 인공지능 알고리즘이 유발하는 차별 방지방안에 관한 연구, 한국거버넌스학회보 제29권 제2호, 2022, 184쪽

성별과 고용 차별의 대표적인 예는 아마존의 직원선발을 위한 인공지능 문제이다. 아마존은 입사지원자들을 평가하고 선별하기 위한 알고리즘을 개발하였으나 인공지능이 남성 지원자를 더 선호하는 결과를 나타내게 된 것이다. 이 인공지능은 '여성'이 언급된 이력서나 여성으로 유추되는 이력서를 저평가하고 남성지원자를 우대하는 결과를 도출하여 아마존은 이 인공지능 개발을 포기하였다.[4]

성별과 금융과 관련하여서는 애플카드 사건이 있다. 애플이 골드만삭스와 함께 애플카드를 출시하였는데 신용한도를 결정을 위한 알고리즘이 남성과 여성을 차별하여 문제가 되었다. 부부가 동일한 자료를 제시하였음에도 불구하고 남성에게 신용한도를 10~20배 더 유리하게 책정한 것이다.[5] 미국은 Equal Credit Opportunity Act에 의하여 성별, 종교, 인종, 피부색 등에 따른 신용차별을 금지하고 있고, 주택임대나 매매와 관련하여서도 Fair Housing Act를 통해 이러한 차별을 금지하고 있음에도 불구하고 이러한 문제가 발생하였다. 다만 사기업의 경우 자체적으로 빅데이터를 이용하여 소비자의 신용평점을 산정할 수 있는데 기존의 데이터의 편향성이 그대로 반영될 가능성이 높다.[6]

2. 인종과 관련한 차별

인종이나 피부색과 관련한 차별도 심각하게 문제된다. 먼저 미국은 COMPAS(Correctional Offender Management Profiling for Alternative Sanctions)라는 범죄예측프로그램을 개발·사용하였다. 이는 미국의 기업인 노스포인트사에 의해 개발된 것인데 알고리즘 설계시에는 인종 등의 변수를 설정하지 않았음에도 불구하고 흑인을 백인보다 2배이상 고위험군으로 판단하여 인종에 의한 차별이라는 문제가 발생되었다.[7]

또한 안면인식과 관련하여 백인보다 흑인이나 유색인종을 식별하는데 있어 차별이 있다. 2021년 페이스북 인공지능이 흑인을 영장류로 분류하는 일이 있었고 2015년에는 구글 포토에서 흑인 사진을 고릴라로 분류하는 사건도 있었다.[8] 실제 안면인식 기술개발에서 백인남성의 데이터가 훨씬 많이 사용되었기 때

4 한애라, 인공지능과 젠더차별, 이화젠더법학 제11권 제3호, 2019, 13쪽

5 애플카드 성차별 논란 휩싸여…"남녀 신용등급 차이 조사중", 한겨레 2019.11.11., https://www.hani.co.kr/arti/economy/finance/916516.html

6 한애라, 인공지능과 젠더차별, 이화젠더법학 제11권 제3호, 2019, 15쪽; 김일우, 고위험 인공지능시스템의 차별에 관한 연구, 서강법률논총 제13권 제1호, 2024, 17쪽

7 이윤아, 윤상오, 인공지능 알고리즘이 유발하는 차별 방지방안에 관한 연구, 한국거버넌스학회보 제29권 제2호, 2022, 187쪽

8 페이스북 AI, 흑인 남성 동영상 '영장류'로 분류 논란, 중앙일보, 2021.09.06., https://www.joongang.co.kr/article/25004708https://www.joongang.co.kr/article/25004708

문에 백인남성에 대한 미인식율이 1%인 반면 흑인여성에 대한 미인식율은 35% 달했다는 조사도 있다.[9]

행정영역에서도 이러한 현상이 발생한 경우가 있다. 영국은 알고리즘을 사용하여 비자발급을 심사했는데, 백인이 신청한 비자는 빠르게 승인되었으나 비백인이나 특정국가 출신자들에게는 심사기간이 오래 걸리거나 비자발급이 허용되지 않는 결과가 나타났다. 이에 영국 내무부는 이 알고리즘 사용을 철회하였다.[10]

의료서비스 제공과 관련하여서도 인종에 대한 차별문제가 나타났다. 미국 민영의료보험사에서 과거 병력 등의 분석을 통해 잠재적인 질병 가능성을 예측하여 질병위험이 높은 사람에게 우선적으로 의료 서비스를 제공하기 위해 AI 시스템을 개발하였으나 실제 질병 위험은 흑인이 높지만 AI는 백인에게 더 우선적인 의료서비스를 제공해야 하는 결론을 제시하였다. 이러한 결과의 원인이 흑인과 백인의 의료비 차이라는 연구도 있다.[11]

3. 사회환경과 관련한 차별

이 외에도 소득이나 지역에 의한 차별이 발생할 수 있다. 영국에서 대학입시에서 인공지능을 통한 시험점수 예측을 통한 입시를 시도하였으나 사립학교 학생들의 점수와 공립학교 학생들의 점수 산정에 차이가 발생하였다. AI 시스템은 전체 학생들 중 40%정도의 시험 점수를 낮게 책정하였는데 이들 대부분이 학비가 저렴한 빈곤지역의 공립학교 학생들이었던 것이다. 이는 지역에 의한 차별도 발생할 수 있음을 보여준다.[12]

4. 기타의 이유로 인한 차별

이뿐 아니라 마이크로소프트사가 사람과 AI가 대화할 수 있도록 개발한 초창기 인공지능인 테이(Tay)의 경우 사용자들이 테이에게 인종차별적, 성차별적, 자극적인 정치적 발언, 혐오표현들을 학습시켜서

---

9 인공지능도 인종차별?…"얼굴인식 소프트에어 백인남성 더 정확", 서울신문, 2018.02.12., https://www.seoul.co.kr/news/international/USA-amrica/2018/02/12/20180212800017

10 이윤아, 윤상오, 인공지능 알고리즘이 유발하는 차별 방지방안에 관한 연구, 한국거버넌스학회보 제29권 제2호, 2022, 187쪽; 편견도 학습하는 AI, 경향신문, 2021.08.11., https://www.khan.co.kr/world/world-general/article/202108112124015

11 AI 의료 알고리즘, "인종 편향성 개선 어려워", 데일리포스트, 2019.12.06., https://www.thedailypost.kr/news/articleView.html?idxno=71803

12 이윤아, 윤상오, 인공지능 알고리즘이 유발하는 차별 방지방안에 관한 연구, 한국거버넌스학회보 제29권 제2호, 2022, 187쪽; 편견도 학습하는 AI, 경향신문, 2021.08.11., https://www.khan.co.kr/world/world-general/article/202108112124015

테이가 차별, 혐오발언을 하게 하여 16시간 만에 서비스를 중단한 경우도 있었고, 국내에서는 SCATTER LAB사에서 열린 주제 대화형 인공지능 (Open-domain Conversational AI)인 '이루다'를 서비스했는데 사용자들이 혐오·차별 표현을 교육시켜 성희롱, 소수자 차별, 혐오 발언 등의 문제로 서비스가 중단된 적이 있다.[13]

III. 인공지능을 통한 차별의 원인

이러한 인공지능으로 인한 여성, 인종, 지역, 장애인, 성소수자에 대한 고용, 금융, 교육, 행정서비스에까지의 차별이 발생하는 이유는 무엇인가?

인공지능과 지금까지의 컴퓨터 기술의 가장 큰 차이점이자 인공지능의 특징은 학습데이터를 입력하여 인공지능을 (머신러닝이든 딥러닝이든 여러 기술을 이용하여) 교육시키는 것과 이를 위한 알고리즘(교육을 위한 알고리즘이든 결과도출을 위한 알고리즘이든)이다. 따라서 불공정성의 이유로 학습데이터가 잘못된 경우와 알고리즘이 잘못 설계된 경우의 두 가지를 상정할 수 있다.

먼저 학습데이터가 잘못된 경우라 함은 데이터의 양이 부족하여 대표할 수 없거나 충분히 학습되지 못한 경우, 데이터의 내용 자체가 편향적이거나 잘못된 데이터가 입력된 경우이다. 전자의 경우는 현실 사회를 적절히 반영할 수 없을 정도의 대표성이 결여된 학습데이터로 AI가 학습하여 다양한 사회의 데이터가 제대로 반영되지 않은 경우라 하겠다. 예를 들어 안면인식 학습데이터가 백인위주로 구성되어 다른 (흑인) 집단의 충분한 데이터가 부족한 경우 대표성이 결여된 학습데이터라 하겠다.

학습데이터의 내용 자체가 편향적인 경우는 세가지 경우로 나눌 수 있는데 개발자의 의도로 편향적인 데이터가 입력되는 경우와 학습데이터를 이용자가 편향적으로 제공한 경우(앞의 이루다나 테이의 경우), 개발자의 의도는 없지만 우리 사회의 현실 자체가 편향적인 경우가 있다.

알고리즘과 관련된 문제는 개발자의 편견이나 편향성이 개입된 경우(고의), 부주의로 인한 경우가 있다. 개발자가 의도를 가지고 인공지능의 편향성, 차별을 유발한 경우로 앞서 예를 들었던 SKT '누구', 애플 시리 등의 AI 개인비서서비스를 예로 들 수 있고, 부주의로 인한 경우는 개발자가 공정성에 대한 고려 없이 알고리즘을 개발하여 개발자 자신의 편견이 (고의는 아니지만) 알고리즘에 반영되어 차별이 발생하는 경우가 있겠다. 이와는 달리 알고리즘 설계에는 문제가 없었으나(편향성이 없이 중립적으로 설계된

13 이윤아, 윤상오, 인공지능 알고리즘이 유발하는 차별 방지방안에 관한 연구, 한국거버넌스학회보 제29권 제2호, 2022, 188쪽; 결국, 잠정 중단된 스캐터랩 AI 챗봇 이루다 사태가 보여준 문제 3가지, AI타임스, 2021.01.12., https://www.aitimes.com/news/articleView.html?idxno=135579

것으로 가정한다면) 결과가 편향적으로 도출되는 경우도 있다. 이 경우는 학습하는 데이터에 문제가 있는 경우라고 하겠다.

결국 인공지능이 편향성을 가지게 되는 것은 학습데이터의 문제일 수도, 알고리즘 설계의 문제일 수도, 이 둘이 합쳐져서 나타나는 결과일 수도 있다.

학습데이터나 알고리즘에 문제가 있는 경우와 관련하여서는 이미 많이 논의되었기 때문에 이 글에서는 이러한 문제 중 학습데이터와 알고리즘에는 문제가 없었으나 결과가 편향적으로 도출되어 차별이 발생하는 경우에 대해 중점적으로 논의하고자 한다.

1. 인공지능의 공정성과 편향성

인공지능은 공정해야 하고 편향적이지 않아야 하는가? 인공지능이 공정할 수 있는가?

인공지능의 공정성과 편향성과 관련하여 이러한 질문을 할 수 있겠지만 어떠한 것이 공정한 것인지, 인공지능이 공정하다는 것은 또는 편향적이지 않다는 것은 어떻게 판단할 것인지에 대해 답을 하기는 어렵다. 그리고 인공지능이 공정할 수 있는지에 대해서도 답하기는 어렵다.

헌법 제11조는 평등에 대해 규정하고 있고 평등권은 중요한 기본권이지만, 이 평등이 우리사회에 절대적으로 지켜지고 있는 것은 아니다. 대표적으로 여성에 대한 또는 피부색이나 민족, 종교로 인한 차별은 전통적으로 존재해 왔으며 이를 해결하기 위해 우리 사회는 적극적평등실현조치(여성할당제) 등 그동안 많은 노력을 기울여 왔다. 그러나 여전히 종교, 성별, 인종, 민족, 성적 지향성, 지역, 정치적 성향, 장애, 나이로 인한 차별은 우리 사회에 존재한다. 우리 사회에 이러한 차별이 존재한다는 것은 현상이고 사실(sein)이다. 하지만 이러한 차별이 잘못된 것이므로 개선되어야 하고 차별이 없어져야 한다는 점은 당위(sollen)이다. 이 점에서 문제가 발생하는데, 인공지능은 기계인데 사람과 유사하게 판단하고 사고할 수 있다고 하여 인공지능에게 sollen을 요구할 수 있는지에 대한 근본적인 문제가 발생한다. 즉 인공지능이 sollen을 학습할 수 있는지, sollen대로 결과(판단)를 도출해 낼 수 있는지, 그러한 결과를 인간이 받아들일 수 있을 것인지의 문제이다.

인공지능의 상용화, 범용화가 이미 실현되고 있고, 인공지능 기술발전의 중요성을 각국이 인정하고 있는 현재 상황에서 인공지능의 사용은 각 분야에서 급속하게 증가할 것이고 우리의 삶에 더욱 밀접하게 관련될 것임은 쉽게 예상할 수 있다. 이러한 상황에서 인공지능의 편향성, 공정성 문제를 지금 논의하지 않으면 우리가 평등을 위해 그동안 노력해왔던 것이 물거품이 될 수 있다. 왜냐하면 인공지능의 사용은 우리 사회의 이러한 편향성과 차별을 반영하는 것에 그치는 것이 아니라 이를 고착화하고 재생산하여 더 확대할 것이기 때문이다.

2. 학습데이터의 편향성 문제

먼저 학습데이터의 편향성 문제를 살펴보자. 학습데이터가 대표성이 부족하거나 개발자에 의해 의도적으로 편향적으로 라벨링된 경우를 규제하는 것은 규범적으로는(기술적 어려움은 차치하더라도) 어려움이 없을 것이다. 다만 문제되는 것이 학습데이터에는 문제가 없었는데(알고리즘도 중립적이라고 가정하고) 결과가 편향적으로 도출된 경우이다. 개발자는 사회에 존재하는, 이미 있는 데이터를 가공하거나 조작하지 않고(소위 Raw data) 그대로 인공지능에게 학습시켰는데, 우리 사회가 편향적이어서 그 데이터에 편향성이 있었고, 인공지능이 이를 학습하여 결과가 편향적으로 도출되는 경우 이를 어떻게 규제할 것인지이다. 인공지능은 가치중립적으로 개발되었고 가공되지 않은 학습데이터를 그대로 학습하여 결론을 제시하는데 이를 공정하지 않고 편향적이기 때문에 평등에 반하여 인공지능 사용정지, 폐기, 개발중지 등의 조치를 할 것인지, 아니면 사회에 존재하는 현상을 그대로 반영한 것이므로 이러한 결과에 정당한 이유가 있다고 볼 것인지 문제이다.

앞서 예를 든 아마존 채용 AI의 경우 아마존에서 10년간 채용관련 자료를 그대로 입력하여 인공지능을 학습시켰고, 여자 지원자를 차별하고자 하는 의도도 없었고 자료에 대한 조작도 없었다고 가정해보자. 단지 과거의 채용관련 자료와 채용에 있어서 남성이 다수를 차지하였기 때문에[14] 인공지능은 10년간의 이러한 데이터를 그대로 학습하여 기존의 비율과 유사한 결론을 제시한 것이라면 이를 인공지능이 여성 지원자를 차별한 것으로 보아야 하는가?

또 다른 예로 COMPAS의 경우도 이와 유사할 수 있다. 그동안 백인보다 흑인의 범죄사실과 재범률이 높았기 때문에 이러한 데이터가 그대로 인공지능에게 학습된 것으로 가정한다면[15] 인공지능이 흑인의 재범확률을 높게 책정하는 것은 그동안의 자료를 바탕으로 한 합리적인 결론이라고 할 수 있을 것이다.[16] 경제적 영역에서도 마찬가지이다. 영국의 대학입시 문제에서도 그 동안의 자료에 의하면 사립학교가 빈곤지역의 공립학교보다 성적이 좋았기 때문에 이를 반영하여 인공지능이 합리적으로 결론을 낸 것이라

14 실제 인력의 60%가 남성, 관리직의 74%가 남성이었다고 한다(James Vincent, Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women, The Verge, October 10, 2018. https://www.theverge.com/2018/10 /10/17958784/ai-recruiting-tool-bias-amazon-report, 손영화, AI 공정성에 관한 연구 – 차별 없는 AI 사회의 실현 –, 한양법학 제34권 제3집, 2023, 280쪽에서 재인용)

15 재범률의 예측 알고리즘 등의 설정이나 전제에 문제가 있다는 부분에 대해 논란이 있다는 점은 차치하고, 만약 이러한 데이터의 입력이 있었다고 가정한다면

16 실제로는 이러한 결과가 검거율 차이에 의한 것이라고도 하고(아마존 채용 AI는 왜 남성을 우대했나, 한국일보, 2021.10.14., https://www.hankookilbo.com/News/Read/A2021101409500001667) 공정성의 기준의 차이로 인한 것이라고 설명하기도 한다.(고학수, 공정한 인공지능의 어려움, 한은소식, 2021. 08.,42쪽 이하) 어떤 이유에서든 사회현상을 반영한 것이라는 점에서는 차이가 없다.

고 해석할 여지도 충분한 것이다. 물론 우리 사회에서 나타나는 이러한 현상이 옳다거나 공정하다거나 편향적이지 않다는 주장을 하려는 것은 아니고, 인공지능에 초점을 맞추었을 때 인공지능이 이러한 결과를 도출하는 것을 편향적인 문제가 있다고 하여 사용중지, 폐기를 할 수 있을 것인지의 문제를 제기하고자 하는 것이다. 기술적인 측면에서는 인공지능이 이러한 결론을 도출한 것은 사용자 또는 개발자의 의도를 잘 반영한 잘 설계되고 학습된 인공지능이라고도 할 수 있을 것이다.

인공지능이 공정하다[17] 혹은 편향적이지 않다는 것은 무엇을 의미하는 것인지에 대한 답은 찾기가 어렵다. 우선 공정하다는 기준을 정하기 어렵고, 결과가 공정하지 않게 나오면 그 인공지능은 공정하지 않은 것이 되는지, 아니면 무엇을 근거로·기준으로 공정, 불공정을 판단할 것인지에 대한 답을 찾기가 어렵기 때문이다.[18] 기계나 기술의 사용에 있어서 가장 공정한 것은 인간의 개입이 가장 적은 것일 수 있다. 인간이 개입하면 할수록 개입한 인간의 편향성이 의도적이든 의도적이지 않든 기술에 개입될 여지가 많아질 수 있기 때문이다. 우리 사회에 이미 편향성이 존재하는 경우 이를 그대로 학습하면 당연히 편향성이 있는 결론을 도출할 수밖에 없다. 인공지능은 기계이기 때문에 sein의 영역에 있다고 볼 수 있다. 우리 사회에서 나타나는 현상, 이를 데이터화 한 자료도 sollen보다는 sein의 영역에 있다고 볼 수 있다. 그렇다면 인공지능에게 sein의 학습데이터를 입력하여 sollen의 결과를 도출하라고 하는 것은 가능한 것인가? 이는 인공지능에게 윤리, 도덕, 법과 같은 당위를 요구할 수 있는가 하는 문제이다. 현재의 기술로는 아직 인공지능이 sollen을 인식하고 학습하는 것은 어려운 문제인 것 같다.

그렇다면 이 문제는 인공지능이 학습하는 데이터의 선별, 조작, 가공을 통해서 해결될 수 있을 것인데, 여기서는 누가 어느 정도로 데이터를 분류하고 조작할 것인지의 문제가 발생한다. 예를 들어 위에서 살펴본 아마존 사건의 경우 학습데이터에 남녀 비율을 동등하게 하여 인공지능을 학습시키는 것을 공정하다고 할 수 있을 것인지의 문제이다. 만약 예를 들어서 실제로는 남성이 가해자인 성범죄가 80~90%임에도 불구하고 인공지능을 위한 학습데이터는 남녀비율을 50:50으로 산술적으로 맞추는 것이 공정한 인공지능이라고 할 것인가, 그렇다면 더 나아가 미국의 COMPAS 문제의 경우에도 흑인의 (재)범죄비율이 백인에 비해 2배로 많다고 가정한다면 이 경우에도 흑인과 백인의 비율을 동일하게 하여 인공지능을 학습시켜 재범죄율을 예측하도록 하는 것이 공정하다고 할 것인지의 문제이다. 외국의 사례 말고 우리 사회의 현실적인 예를 들어보면, 네이버에 보수와 진보언론의 노출 비율이 7:3이면 우리는 이를 보수편향의 불공정한 편향성을 가진 서비스라고 할 수 있을 것이다. 적어도 보수와 진보언론 노출비율이 5:5 가까이 되어야 공정한 서비스라고 볼 여지가 있기 때문이다. 그러나 그 사회에 보수와 진보언론의 수와 기사의 비율이 7:3이라면

17 사실 공정이라는 개념이 다의적이고 각 사회마다 다르게 해석될 수 있는 부분이어서 공정하다는 정의 자체를 내리기가 어려운 부분도 있다.

18 구체적으로 어떤 공정성 기준을 적용하여 알고리즘을 평가하는지에 따라, 인공지능이 활용되는 다양한 맥락에 따라 공정성에 대한 다른(때로는 반대의) 결론이 도출될 수 있다(고학수, 공정한 인공지능의 어려움, 한은소식, 2021. 08. 43쪽)

(신문은 경향성(Tendenz)이 보호되기 때문에) 이를 있는 그대로 7:3의 비율로 노출되게 하는 것이 공정한 것인가 아니면 인위적으로 5:5로 맞추어 이용자들에게 제공하는 것이 공정한 것인가? 이에 대해 답하기는 어렵다. 또한 만약 이러한 산술적·양적·결과적 공정성에 입각하여 학습한 인공지능을 사용하여 도출된 결과는 우리가 수용하여 이용할 수 있을 정도의 정확성을 가지는지, 혹은 개발 목적에 부합하는 것이라고 할 수 있을지도 의문이다. 더 나아가 우리는, 우리 사회는 그렇지 않음에도 불구하고 인공지능에게 너무 강한 윤리와 공정성을 요구하는 것은 아닌지하는 우려도 생각해볼 수 있을 것이다. 또한 학습데이터를 선별, 가공함으로써 새롭게 발생할 수 있는 편향성을 어떻게 해소할 것인지도 같이 논의되어야 한다.

3. 알고리즘의 편향성 문제

학습데이터를 가공하는 것에 대한 우려가 있다면 알고리즘을 설계할 때 이러한 부분을 고려하도록 설계하는 방법을 모색해 볼 수 있다. 학습데이터가 sein의 영역에 가깝다면 알고리즘의 설계는 인간이 하는 것이기에 sollen을 가미할 수 있는 영역이기에 설계에 있어서 이러한 점을 고려할 것을 명령하거나 요구할 수는 있을 것이다. 다만 알고리즘으로 인한 편향성은 더욱 위험한 결론을 발생시킬 수 있다는 점-인간의 개입으로 인한 인위적인 결과의 조정이기 때문에-이 우려되는 부분이다.

알고리즘으로 인한 편향성의 이유로 세 가지 경우를 상정할 수 있는데, 첫째는 알고리즘 설계단계에서 고의로 편향적으로 설계한 경우, 둘째는 고의는 없었지만 개발자의 무의식적인 편향성이 알고리즘 설계에 투영된 경우를 상정할 수 있다.[19] 세 번째는 두 번째 이유와 구분하기는 어려우나 이러한 사회적 차별 문제를 심각하게 고려하지 않고(기존의 편향성을 조정하지 않고) 설계한 경우를 들 수 있겠다.[20](이는 학습데이터와 연결되어 편향성을 나타내게 될 것이다).

알고리즘을 고의로 편향적으로 설계하여 인공지능 차별문제가 발생한 경우 학습데이터의 편향성으로 인한 문제보다 심각한 결과를 발생시킨다. 우리사회의 편향성을 인공지능이 그대로 반영하는 경우 이는 어느정도 용인될 수 있는 여지가 있으나 알고리즘의 편향성으로 인한 차별의 문제는 인위적인 차별이기 때문이다. 예를 들어 금융과 관련하여 인공지능을 통하여 대출심사(금리산정) 또는 신용점수를 산정할 때 기존의 위험요소에 대한 데이터를 학습데이터로 하여 인공지능을 개발하는 것과 이에 더하여 여자인지, 흑인인지에 따른 변수를 부정적으로 고려할 것을 추가하여 알고리즘을 설계하는 것은 본질적으로 다른 문제일 수 있다. 이러한 차별은 기존의 법으로도 규제하거나 금지하는 것이 가능할 것으로 보인다.

---

19 이준일, 인공지능과 헌법, 헌법학연구 제28권 제2호, 2022, 365쪽; 박도현, 인간 편향성과 인공지능의 교차, 서울대학교 법학 제63권 제1호, 2022, 151쪽(이를 '의식적인' 편향성과 '암묵적인' 편향성으로 구분한다)

20 이를 알고리즘의 특성인 불투명성, 예측불가능성으로 인한 것으로 보기도 한다,(김성용, 정관영, 인공지능의 개인정보 자동화 처리가 야기하는 차별 문제에 관한 연구, 서울대학교 법학 제60권 제2호, 2019, 326쪽)

두 번째의 경우는 이러한 편향성을 발견하기 어렵다는 문제점을 가진다. 이는 알고리즘을 작동시키는 연산과정의 비가시성[21]과 불투명성(설명의 어려움)에 기인한다. 알고리즘의 불투명성은 개발자들이 영업비밀로 공개하기를 꺼려하기 때문이기도 하고 전문가가 아닌 자들이 작동방식을 이해할 만큼의 전문적인 지식이 부족하거나 전문가조차도 그 작동방식을 이해하기 어렵기 때문이기도 하다.[22] 예를 들어 네이버는 이미 정치분야 뉴스 배치의 편향성 문제를 지적받았고, 이러한 사회적 비판에 대해 2018년 네이버는 사람을 통해 뉴스 편집을 하지 않고 인공지능에 뉴스 편집을 맡기고 인간은 기술을 지원하는 역할만 한다고 선언하였다. 하지만 네이버의 뉴스배치는 보수적 편향성을 여전히 드러낸다는 비판이 많기 때문에 뉴스배치 알고리즘에 대한 의심이 제기되지만 네이버는 뉴스 편집 알고리즘 기술이 영업 비밀이라는 이유로 공개는 하지 않고 있다.[23]

세 번째의 경우는 알고리즘을 학습데이터로 인해 나타날 수 있는 편향성을 제거하거나 줄이도록 설계하거나 변형하도록 규범적으로 강제할 수 있을 것인가 하는 문제를 발생시킴은 물론 알고리즘을 인위적으로 조정함으로써 인공지능의 편향성 문제를 해결할 수도 있으나 알고리즘 편향성으로 인한 새로운 평등문제를 야기할 수도 있다는 문제점을 가진다. 알고리즘 설계나 조작을 규제하는 것은 쉬운 일은 아니다. 일단 인공지능 등을 설계하고 학습하는 알고리즘은 개발자 또는 개발업체의 많은 비용과 시간, 인력을 투입하여 개발한 것이므로 영업비밀에 속하여 이를 공개하는 것을 거부할 확률이 매우 높고, 이를 공개하라고 명령하거나 개발과정을 설명하거나 밝히라고 하기도 매우 어렵다. 이를 설명의무, 설명가능성, 설명요구권, 알고리즘의 투명성 등으로 설명하지만 이것이 사실상 실현되는 것을 어렵다고 보여진다.

예를 들어 2022년 서울고등법원에서는 네이버가 알고리즘을 조작하여 자사 서비스를 우대하여 공정거래위원회로부터 시정조치와 과징금을 부과받는 사건에서 네이버 패소판결을 한 바 있는데[24] 이에 대해 네이버는 알고리즘 변경은 검색엔진에서 일상적인 일이라며 검색 알고리즘 조정은 소비자의 효용 증진을 위한 것이었다고 주장[25]하면서 대법원에 상고한 상태이다. 또한 쿠팡은 쿠팡에게 알고리즘을 조작하여 자체 브랜드(PB)를 부당하게 우대했다는 이유로 공정거래위원회로부터 과징금 1400억원을 부과받았

---

21 정원섭, 인공지능 알고리즘의 편향성과 공정성, 인간.환경.미래 제25호, 2020, 65쪽

22 원상철, 인공지능의 윤리와 법의 접점. 법이론실무연구 제12권 제2호, 2024, 213쪽; 김지연, 인공지능(AI)의 윤리적 지위: 인간과 비인간 사이에서 어울리기, 사회와 이론 46, 2023

23 드러난 보수편향 네이버 뉴스 편집 AI알고리즘...공정성 논란 불붙나, AI타임스, 2021.03.08., https://www.aitimes.com/news/articleView.html?idxno=137148

24 서울고등법원 2022. 12. 14. 선고 2021누36129

25 [판결] "'비교쇼핑 검색 알고리즘 조작 혐의' 네이버에 266억 과징금 부과 정당", 법률신문, 2022.12.15., https://www.lawtimes.co.kr/news/183827; [기획]네이버, 공정위와 '검색 알고리즘 공정성' 두고 법정공방, 매일일보, 2023.02.19., https://www.m-i.kr/news/articleView.html?idxno=989013

으나 쿠팡은 정상적인 알고리즘 조작이라고 주장하고 있다.[26]

이러한 예는 알고리즘 조작의 위험성과 그 규제의 어려움을 드러내는 중요한 사건이라고 할 수 있다. 위의 사건은 단순히 알고리즘 자체의 공정성에 대한 문제가 아니라 기업 내부자의 경제적 목적을 위한 인위적 개입이 문제가 된 것이지만, 기업들은 모두 성능 향상을 위한 불가피한 조치인 정상적 조치라고 주장하고 있다. 또한 기업들은 자신들의 알고리즘을 저작권, 영업비밀 등의 이유로 공개하기를 꺼려하기 때문에 어떠한 알고리즘으로 그러한 결과가 도출되었는지 알기가 어렵다. 이 사건은 관련 법률이 있고, 조사와 증거가 있었기에 어느정도 밝혀진 것이지만, 성별, 인종 등에 의한 차별과 관련된 알고리즘의 조작은 편향성을 발견하기도 이를 증명하기도 어렵다는 문제가 있다.

마지막으로 인공지능의 공정성을 위해 학습데이터를 적극적으로 조작하고 알고리즘을 설계하거나 조작하는데 있어 기존의 편향성을 조정·완화하기 위해 적극적으로 개입하도록 하는 것은 어떻게 정당화될 수 있을 것인가? 또한 적극적으로 이러한 조치를 취해야 한다면 어느 수준까지 평등하여야 공정한 인공지능이라고 인정될 수 있을 것인지가 논의되어야 한다.

우선적으로 이러한 논의는 규범적으로는 적극적 평등실현조치(affirmative action)에서의 논의와 유사성을 가진다. 적극적 평등실현조치는 역사적으로 차별을 받아온 집단에게 그 동안의 차별을 보상해 주기 위해 그 집단을 우대하는 조치를 취하는 국가의 잠정적 우대조치를 의미한다. 대표적인 예로 여성할당제를 들 수 있다.[27] 현실은 그렇지 아니하지만 이러한 현실을 개선하기 위해 학습데이터를 입력할 때 그동안의 소수자들에게 더 우선권을 줄 것, 가치중립적인 알고리즘을 개발하는 것이 아니라 우리사회에 차별받는 자들을 우선할 수 있는(또는 차별받지 않는 집단과 유사하게 취급될 수 있는) 가치지향적 알고리즘을 개발할 것을 요구하는 것이기 때문이다.

그리고 인공지능에서의 공정성을 위한 조치를 산술적 평균으로 해야 하는지에 대한 논의도 필요하다. 예를 들어 흑인 또는 백인의, 남성 또는 여성의 범죄율이 높음에도 불구하고 학습데이터는 인종성별의 비율을 동일하게 하도록 할 것인지, 알고리즘을 조작하여 결론이 인종성별의 비율이 동일하게 도출되도록 할 것인지의 문제이다. 고용에 있어서도 마찬가지이고, 정치적 성향에 있어서도 마찬가지이다. 앞서 예를 든 뉴스포털의 경우 뉴스노출을 진보와 보수 5:5로 하도록 할 것인지가 문제될 것이다. 현실의 국민의 의견은 그렇지 않음에도 불구하고 보여지는 것에 산술적 평균이 요청된다면 이는 오히려 인공지능에

26 "쿠팡, 알고리즘 조작으로 자사상품 1등 만들었다…임직원, 후기 7만개 작성", 동아일보, 2024.06.13. https://www.donga.com/news/Economy/article/all/20240613/125411557/2

27 그러나 이러한 적극적 평등실현조치는 다시 다른 집단의 불평등을 야기하므로 이를 어느정도까지 인정할 것인지, 어느정도가 헌법적으로 허용될 수 있을 것인지는 다시 판단하여야 한다.

의해 인간의 의사형성이 왜곡되거나 조종되는 것(manipulate)이라고 볼 수도 있는 것이다.[28]

IV. 인공지능의 공정성 확보를 위한 방안

앞서 설명한 인공지능의 편향성 문제를 해결하기 위한 가장 좋은 방안은 우리 사회의 편향성을 없애는 것이고, 그 다음 좋은 방안은 알고리즘 개발자가 편향성이 없거나 편향성을 적극적으로 해결하고자 하는 것이다. 결국 지금까지의 인공지능의 윤리에 대한 논의는 인공지능 기계 자체에 윤리나 sollen을 교육하기보다는 개발자에게 윤리를 요구하는 것에 가깝다.

인공지능의 공정성을 확보하기 위해서는 학습데이터의 입력에서부터 알고리즘의 설계에 이르기까지 여러 단계에서의 규제방법이 제시되고 있다. 대표적으로는 학습데이터와 관련한 데이터 거버넌스 문제와 알고리즘과 관련한 설명의무, 설명요구권이 논의되고 있다. 아래에서는 여러 국가에서의 인공지능 공정성 확보를 위한 방안을 살펴본다.

1. EU

EU에서는 AI법을 제정하여 법에서 인공지능의 차별 문제를 규제하고 있다. 이를 위해 투명성을 강조하면서 인종, 정치적 의견, 노동조합 가입, 종교적 또는 철학적 신념, 성생활 또는 성적지향을 추론하기 위하여 생체인식데이터를 기반으로 자연인을 개별적으로 분류하는 생체인식분류시스템(biometric categorisation system)의 사용을 금지한다. 또한 개인이나 집단의 나이, 장애나 특정한 사회적·경제적 상황에 의한 취약성을 악용하는 인공지능시스템의 사용, 사회적 점수를 통하여 자연인이나 집단의 사회적 행동 또는 알려지거나 추론되거나 예상되는 성격이나 특성에 기초하여 특정한 기간 동안 자연인이나 집단을 평가하거나 분류하기 위한 인공지능시스템의 사용을 금지한다. (제5조)

뿐만아니라 이러한 고위험 인공지능을 규제하기 위해 데이터의 수집과 라벨링(labelling), 필요한 데이터의 가용성, 수량 및 적합성 평가 등을 실시하며, 유럽연합 법령에 따라 금지된 차별을 초래하는 편향 가능성을 고려한 조사, 확인된 편향 가능성을 감지, 방지 및 완화하기 위하여 적절한 조치를 시행할 것을 명령하고 있다. 특히 제10조에 데이터와 데이터거버넌스 규정을 두어 학습데이터 품질기준을 규율하고자 한다.

---

28 이와 관련하여서는 문의빈, 사상의 자유의 재조명 — 인공지능에 의한 의사형성과정의 조종 가능성을 중심으로 —, 공법연구 제52집 제3호, 2024 참고

2. 미국

미국의 2022 국가인공지능계획법에 의하면 국가인공지능 자문위원회(National Artificial Intelligence Advisory Committee)를 구성하여 이러한 문제에 대응하고자 하고 있다. 또한 2022년 10월 AI 권리장전(AI Bill of Rights)[29]이 발표되었고, 2023. 10. 30. 바이든 정부는 안전과 보안이 보장되며 신뢰할 수 있는 인공지능에 관한 행정 명령(Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)[30]을 발표하였다.[31] 이 행정명령은 인공지능이 가져올 다양한 이점과 함께 활용에 따른 위험을 내포하고 있으므로 인공지능 개발·사용시 준수해야 할 행정부의 8가지 원칙[32]을 제시하고, 행정명령을 위해 관련정책을 입안·이행하고 정부의 활동을 조정하기 위해 백악관에 AI위원회를 설치하도록 했다.[33] 민간차원에서는 2022 알고리즘 책임법 (Algorithmic Accountability Act of 2022)을 통해 알고리즘 활용으로 인한 오류나 편향성 등으로 인한 문제를 예방하고 규제하고자 한다. 이 법은 이를 위해 사전 영향평가라는 안전장치를 마련하고, 연방통상위원회를 통하여 이러한 영향평가의 실효성을 보장하고자 한다.[34] 미국은 한편으로는 위원회를 설치하여 이러한 문제에 유연하게 대응하고자 하고, 다른 한편으로는 공공과 민간에 적용되는 법률을 달리 두어 규범적으로 강제하고자 한다.

3. OECD

OECD는 2019년 인공지능에 관한 이사회 권고안(Recommendation of the Council on Artificial Intelligence)을 발표한 바 있고, 2024년 5월에는 업데이트된 인공지능 권고안을 발표하였다. 이에 의하

---

29 https://www.whitehouse.gov/ostp/ai-bill-of-rights/

30 행정명령 14110호

31 https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

32 인공지능기술의 안전과 보안의 보장(Ensuring the Safety and Security of AI Technology), 혁신과 경쟁의 촉진(Promoting Innovation and Competition), 노동자에 대한 지원(Supporting Workers), 평등과 시민권의 증진(Advancing Equity and Civil Rights), 소비자·환자 등의 보호(Protecting Consumers, Patients, Passengers, and Students.), 사생활(개인정보) 보호(Protecting Privacy), 인공지능 사용에 관한 연방정부의 노력(Advancing Federal Government Use of AI), 전지구 차원에서의 미국의 리더쉽 강화(Strengthening American Leadership Abroad) 등

33 법제처 미래법제혁신기획단, 인공지능(AI) 관련 국내외 법제 동향, 2024, Legislation Newsletter, 31쪽

34 알고리즘 책임법에 나타난 평등보호 관련 규정에 대한 더 구체적인 내용은 김일우, 고위험 인공지능시스템의 차별에 관한 연구, 서강법률논총 제13권 제1호, 2024, 33쪽 참고

면 인공지능 시스템이 고용, 금융, 의료 등 다양한 영역에서 불합리한 차별을 초래하지 않도록 해야 함을 강조하면서 인공지능이 의도적 편견뿐만 아니라 의도적이지 않은 편향으로도 부당한 결과를 낳을 수 있음을 지적하고 인공지능 개발자들이 편향성 및 불공정성의 위험을 사전에 분석하고 시정조치를 마련해야 한다고 권고한다. 또한 인공지능의 투명성과 설명가능성(Transparency and explainability)을 강조하여 인공지능 개발자와 운영자는 책임 있는 공개(responsible disclosure)를 실천하고, 인공지능의 결정과 관련된 자들이 결과에 이의를 제기하거나 설명을 요구할 권리를 보장할 것을 권고한다.

V. 나가며

이미 세계 각국에서는 인공지능의 편향성으로 인한 부작용, 불평등 문제에 대해 인식하고 이를 개선하기 위해 경성적이든 연성적이든 규범을 통하여 해결하고자 하는 움직임을 보이고 있다.

하지만 앞서 살펴본 바와 같이 규범의 영역에서 해결이 가능한 편향성 문제가 있는 반면, 숨겨져서 우리 사회의 불평등을 그대로 재생산하거나 불평등을 확대하는 편향성 문제가 있다.

이미 문제가 되어 많은 논의가 있었던 편향성 문제는 어느 정도 해결의 방법을 찾은 것 같다. 데이터의 대표성이 부족하다거나 학습데이터가 처음부터 잘못 입력된 등의 사례는 오히려 해결이 쉽다.

하지만 학습데이터에도 문제가 없고, 인공지능의 알고리즘 자체에도 문제가 없지만 인공지능의 결과는 편향적으로, 불평등하게 도출되는 경우 이를 어떻게 해결할 것인지에 대해서는 아직 논의가 부족한 것으로 보인다. 대표적으로 우리사회의 편향성(여성에 대한, 인종에 대한 차별)을 인공지능이 그대로 학습하여 편향적인 결론을 도출하는 경우 이를 공정하지 않은 인공지능이라고 할 수 있을 것인지의 문제이다.

실제로 학습데이터에도 문제가 없고, 알고리즘에도 문제가 없지만 인공지능이 편향적, 불평등한 결론을 도출하는 경우 이 인공지능의 사용을 강제로 금지할 수 있을 것인지, 그 보상이나 개발비용의 보전은 어떻게 할 것인지 등의 법적인 문제가 될 수도 있을 것이다.

인공지능이 sollen을 학습하여 우리 사회의 편향적인 부분을 스스로 조정하여 윤리적으로 당위적으로 공정한 결론을 도출하라고 명령할 수 있을 것인가? 인공지능이 sollen을 학습은 할 수 있을 것인가? 어떠한 결론이 공정하다, 편향적이지 않다라고 판단할 수 있을 것인가? 학습데이터가 혹은 알고리즘이 어느 정도로 조정되어야 편향적이지 않은 인공지능이라고 할 수 있을 것인가?

아직 이러한 문제에 대하여 우리 사회에서 선뜻 또는 공통된 답을 하기는 어려운 것 같다. 우리는 아직 인공지능의 공정성, 편향성, 불평등과 관련하여 더 많은 논의와 고민과 사회적 합의가 필요하다.

하지만 중요한 것은 인공지능이 조금 더 상용화되기 전에 이러한 부분에 대한 고민을 하여 규범적(강제적)으로 혹은 자율규제의 방식으로라도 어느 정도의 가이드라인은 도출되어야 할 것이다. 왜냐하면 우리 사회의 편향성을 인공지능이 학습하기 시작한다면 sollen을 모르는 인공지능으로 인하여 우리 사회의 불평등이 재생산됨은 물론 확대될 것이 분명하고, 이제까지 노력해왔던 우리 사회의 불평등 해소를 위한 노력이 물거품이 될 우려가 있기 때문이다.

AI and the Reproduction of Inequality

LEE Kwon il | Associate Professor, School of Law, Kyungpook National University

Ⅰ. Introduction

Terms such as the Fourth Industrial Revolution, the intelligent information society, big data, artificial intelligence (AI), smart cities, and smart farms are no longer unfamiliar; they have become part of the everyday vocabulary of the general public. Social changes driven by advances in IT (information technology) are accelerating, and our daily lives are being transformed accordingly. Among these developments, the advancement of artificial intelligence has recently become the most widespread and widely discussed.

The 2016 match between AlphaGo and professional Go player Lee Sedol left a profound impact, demonstrating the rapid advancement of AI technologies. By late 2022, with the launch of widely accessible generative AI services such as ChatGPT (Generative Pre-trained Transformer), the public witnessed their rapid evolution in real-world markets. These developments have sparked intense debate on the commercialization of AI, highlighting both its potential benefits and the challenges it entails. In particular, as AI has recently come into wide commercial use and AI-powered devices have received strong responses in the market, global tech companies such as Google and Microsoft, as well as companies like Samsung and Apple, along with governments around the world, are investing substantial human and financial resources in AI development and are staking everything on gaining a competitive edge. This widespread adoption of artificial intelligence marks an important turning point: AI technology has moved beyond the stage of being discussed solely among scientists, developers, or other experts, or being offered merely as beta services, and has now become deeply integrated into our daily lives.

The advancement of artificial intelligence brings with it a belief that it will offer us great convenience, but at the same time it also generates concerns and fears about the unpredictability of its outcomes. In addition, artificial intelligence has been criticized for its opacity, as the processes through which the technology operates (algorithms) are often difficult to know, inaccessible, or in some cases deliberately withheld from disclosure. This paper seeks

to address two major issues concerning artificial intelligence: bias (fairness) in AI and the AI's reproduction of inequality. To this end, this paper will examine cases of discrimination arising from AI bias (Section II), explore the cause of such phenomena (Section III), and discuss possible solutions (Section IV).

II. Cases of Discrimination Caused by Algorithmic Bias

Even though AI may exhibit human-like cognition, it remains a machine rather than a human being. And until now, we have held the belief that machines are value-neutral; therefore, it has been assumed that, as long as AI is a machine, it would be more fair and less biased than humans. This belief is well demonstrated by cases such as Naver's AI-enabled news editing[1] without human intervention, or the introduction of AI-based job interviews.

When examining the outcomes of AI development and deployment, it becomes clear that issues of equality, discrimination, and hate stemming from algorithmic bias constitute a significant part of AI's negative side effects. While many of the examples we cite are from other countries—and one might argue that the patterns of discrimination differ from Korean context, since issues of discrimination based on race and religion are more prominently highlighted overseas—Korea is by no means immune from such problems. In Korea, serious concerns also arise from discrimination based on gender, sexual orientation, age, educational background, and region, as well as from fairness issues linked to political polarization.

This paper starts by examining representative examples and forms of discrimination caused by artificial intelligence to analyze patterns of discrimination and explore possible solutions.

A. Discrimination Based on Gender

Let us first look at discrimination related to gender. Gender-related issues include gender assigned to specific AI tool and the AI's reinforcement of gender roles. For example, gender assigned to AI virtual assistants—through names, voices, and other features—is often female. AI virtual assistants such as SKT's NUGU, KT's Giga Genie, Naver's Friends, Kakao's Mini, Amazon's

1 Naver Gives Up News on Its Mobile Front Page," Media Today, May 9, 2018, available at: https://www.mediatoday.co.kr/news/articleView.html?idxno=142622. Concerns were also raised regarding fairness in this case.

Alexa, Microsoft's Cortana, and Apple's Siri not only default to female voices, but in many cases also did not allow users to change this setting. This practice not only reflects existing issues of gender roles in our society but also raises concerns that it may reinforce gender bias.[2]

In addition, Google's automatic translation service has reproduced existing gender role practices by translating occupations such as engineers and doctors as male, while assuming jobs like nurses to be female.[3] A representative case of gender and employment discrimination is Amazon's AI system for recruitment. The AI downgraded résumés that mentioned "women" or could be inferred as belonging to female applicants, while favoring male candidates, which ultimately led Amazon to abandon the development of the system.[4]

An example that illustrates the AI's bias in gender and finance is the Apple Card case. Apple, together with Goldman Sachs, launched the Apple Card, but the algorithm used to determine credit limits discriminated between men and women, leading to controversy. Even when spouses submitted identical information, the algorithm set credit limits 10 to 20 times higher for men than for women.[5] In the United States, the Equal Credit Opportunity Act prohibits credit discrimination based on gender, religion, race, or color, and the Fair Housing Act likewise prohibits such discrimination in housing rentals and sales. Nevertheless, such discrimination persists. However, for private companies, credit scores for consumers may be calculated using big data, and there is a high likelihood that the biases embedded in existing data will be directly reflected.[6]

---

2 Ae Ra Han, Artificial Intelligence and Gender Discrimination, Ewha Journal of Gender and Law, Vol. 11, No. 3, 2019; Yuson Heo, An Analysis of Cases of Gender Discrimination in Artificial Intelligence, Journal of Ethics Education, Vol. 71, 2024, p. 461 ff.

3 Yoonah Lee and Sangoh Yoon, A Study on Measures to Prevent Discrimination Caused by AI Algorithms, Korean Journal of Governance, Vol. 29, No. 2, 2022, p. 184.

4 Ae Ra Han, Artificial Intelligence and Gender Discrimination, Ewha Journal of Gender and Law, Vol. 11, No. 3, 2019, p. 13.

5 "Apple Card Engulfed in Gender Discrimination Controversy… 'Investigating Differences in Credit Limits Between Men and Women'," The Hankyoreh, November 11, 2019, available at: https://www.hani.co.kr/arti/economy/finance/916516.html.

6 Aera Han, Artificial Intelligence and Gender Discrimination, Ewha Journal of Gender and Law, Vol. 11, No. 3, 2019, p. 15; Ilwoo Kim, A Study on Discrimination in High-Risk AI Systems, Sogang Law Review, Vol. 13, No. 1, 2024, p. 17.

B. Discrimination Based on Race

Discrimination based on race or skin color is also a serious concern. To begin with, the United States developed and used a crime prediction program called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). This program was developed by the U.S. company Northpointe. Although variables such as race were not included in the algorithm's design, it nevertheless classified Black defendants as high-risk at more than twice the rate of White defendants, raising concerns of racial discrimination.[7] Facial recognition is not immune to discrimination, where Black people and people of color are identified less accurately than White individuals. In 2021, Facebook's AI classified Black people as "primates," and in 2015, Google Photos categorized photos of Black individuals as "gorillas."[8] In fact, studies have shown that because facial recognition technologies were developed using disproportionately more data from White men, the misidentification rate for White men was only about 1%, whereas for Black women it reached as high as 35%.[9]

Such cases have also occurred in the administrative sector. In the United Kingdom, algorithms were used to review visa applications: while applications from White applicants were approved more quickly, those from non-White applicants or individuals from certain countries faced longer processing times or were denied altogether. As a result, the UK Home Office withdrew the use of this algorithm. [10]

Racial discrimination has also appeared in the provision of medical services. A U.S. private health insurance company developed an AI system to analyze medical histories and other data in order to predict potential disease risks and to prioritize medical services for those at

---

7 Yoonah Lee and Sangoh Yoon, A Study on Measures to Prevent Discrimination Caused by AI Algorithms, Korean Journal of Governance, Vol. 29, No. 2, 2022, p. 187.

8 "Facebook AI Sparks Controversy by Classifying Video of Black Men as 'Primates'," JoongAng Ilbo, September 6, 2021, available at: https://www.joongang.co.kr/article/25004708.

9 AI and Racial Discrimination?… 'Facial Recognition Software More Accurate for White Men'," Seoul Shinmun, February 12, 2018, available at: https://www.seoul.co.kr/news/international/USA-america/2018/02/12/20180212800017

10 Yoonah Lee and Sangoh Yoon, A Study on Measures to Prevent Discrimination Caused by AI Algorithms, Korean Journal of Governance, Vol. 29, No. 2, 2022, p. 187; "AI that Learns Biases," Kyunghyang Shinmun, August 11, 2021, available at: https://www.khan.co.kr/world/world-general/article/202108112124015.

higher risk. However, although Black patients actually faced higher health risks, the AI system concluded that White patients should be given higher priority for medical services. Some studies suggest that this outcome was driven by differences in medical expenditures between Black and White patients. [11]

C. Discrimination Related to Social and Environmental Factors

Discrimination may also arise based on income or region. In the United Kingdom, an attempt was made to use AI to predict exam scores for university admissions, but differences emerged between the scores calculated for students from private schools and those from public schools. The AI system downgraded the exam scores of about 40% of all students, most of whom were from public schools in low-income areas with lower tuition fees. This demonstrates that discrimination can also occur on the basis of region.[12]

D. Discrimination for Other Grounds

In addition, Microsoft's early conversational AI, Tay, which was designed to interact with people, had to be shut down after just 16 hours when users trained it with racist, sexist, inflammatory political statements, and hate speech, leading the system itself to produce discriminatory and hateful remarks. Similarly, in Korea, SCATTER LAB launched Iruda, an open-domain conversational AI, but the service was discontinued after users taught it discriminatory and hateful expressions, which resulted in problems such as sexual harassment, discrimination against minorities, and hate speech.[13]

---

11 "AI Medical Algorithms: 'Difficult to Improve Racial Bias'," The Daily Post, December 6, 2019, available at: https://www.thedailypost.kr/news/articleView.html?idxno=71803.

12 Yoonah Lee and Sangoh Yoon, A Study on Measures to Prevent Discrimination Caused by AI Algorithms, Korean Journal of Governance, Vol. 29, No. 2, 2022, p. 187; "AI that Learns Biases," Kyunghyang Shinmun, August 11, 2021, available at: https://www.khan.co.kr/world/world-general/article/202108112124015.

13 Yoonah Lee and Sangoh Yoon, A Study on Measures to Prevent Discrimination Caused by AI Algorithms, Korean Journal of Governance, Vol. 29, No. 2, 2022, p. 188; "Three Problems Revealed by the Suspension of Scatter Lab's AI Chatbot Iruda," AI Times, January 12, 2021, available at: https://www.aitimes.com/news/articleView.html?idxno=135579.

Ⅲ. Sources of AI-Driven Discrimination

What explains the discrimination caused by artificial intelligence—discrimination that extends to women, racial and regional minority groups, persons with disabilities, and sexual minorities in areas such as employment, finance, education, and administrative services?

The most significant difference between artificial intelligence and conventional computer technologies—and the defining feature of AI—is that it is trained on datasets using various techniques such as machine learning and deep learning, and that this process relies on algorithms (both those used for training and those used for producing results). Accordingly, two main sources of unfairness can be identified: cases where the training data are flawed, and cases where the algorithms are poorly designed.

First, flawed training data may refer to situations where the dataset is too small or unrepresentative, preventing AI from adequately reflecting the diversity of real-world society, or where the content of the data itself is biased or erroneous. For example, when facial recognition training data are composed primarily of White individuals, with insufficient data from other groups such as Black individuals, the dataset can be said to lack representativeness.

Training data bias can be divided into three categories: cases where biased data are intentionally entered by the developer; cases where users supply biased data (as in the earlier examples of Iruda and Tay); and cases where there is no such intent, but the data nonetheless reflect the inherent biases of society itself.

Problems related to algorithms may arise either from the developer's deliberate incorporation of prejudice or bias, or from negligence. Examples of developers' intentional design of AI in ways that foster bias or discrimination can be found in AI personal assistant services such as SKT's NUGU and Apple's Siri, as mentioned earlier. By contrast, cases of negligence occur when developers create algorithms without considering fairness, thereby allowing their own prejudices—though not deliberate—to be reflected in the system and result in discriminatory outcomes. Meanwhile, some algorithms are properly designed—assuming they were created in a neutral, unbiased manner —yet the results still turn out to be biased. In such instances, the problem lies in the training data.

Ultimately, the bias in artificial intelligence may arise from problems in the training data, from flaws in algorithm design, or from the combined effect of both.

Since cases involving problems with training data or algorithms have already been extensively discussed, this paper will focus instead on situations where, even in the absence of flaws in the data or the algorithm, the results nevertheless turn out to be biased and lead to discrimination.

A. Fairness and Bias in Artificial Intelligence

Should artificial intelligence be fair and unbiased? Can AI truly be fair?

These are the kinds of questions that may be raised in relation to fairness and bias in AI, yet it is difficult to answer what fairness actually means, how one can judge that AI is fair or unbiased, and whether AI can in fact be fair at all.

Article 11 of the Constitution of the Republic of Korea provides for equality, and the right to equality is a fundamental constitutional right. However, this principle of equality is not absolutely upheld in Korean society. Discrimination against women, or on the basis of skin color, ethnicity, and religion, has long existed. To address this, Korean society has made considerable efforts, including measures such as affirmative action (e.g., gender quota systems). However, discrimination based on religion, gender, race, ethnicity, sexual orientation, region, political orientation, disability, and age continues to exist in Korean society. The existence of such discrimination in our society is a reality and a fact (Sein). However, the recognition that such discrimination is wrong and must be remedied, and that it should ultimately be eliminated, belongs to the realm of obligation (Sollen). This raises a fundamental problem: although AI is a machine, and while it is said to be capable of reasoning and making judgments similar to humans, a core question arises as to whether Sollen can be demanded of AI. In other words, can AI learn Sollen? Can it produce outcomes (judgments) in accordance with Sollen? And can such outcomes be accepted by human beings?

With the commercialization and widespread use of artificial intelligence already underway, and with countries around the world recognizing the importance of AI development, it is easy to anticipate that the use of AI will rapidly expand across various sectors and become ever more closely intertwined with our daily lives. In this context, if we fail to address the issues of bias and fairness in artificial intelligence now, the efforts we have made thus far to advance equality may come to nothing. This is because the use of AI does not merely reflect existing biases and discrimination in our society, but rather entrenches and reproduces them, thereby amplifying their effects.

B. Biased Training Data

Let us examine biased training data. From a normative perspective (setting aside technical difficulties), regulating cases where training data lack representativeness or are intentionally labeled in a biased way by developers would pose little difficulty. The more challenging problem arises when neither the training data nor the algorithm is flawed (assuming the algorithm is neutral), yet the outcomes are biased. In such cases, developers may have trained AI on unprocessed "raw data" without alteration or manipulation, but because society itself is biased, the data inherently contained biases. When AI learns from such data and produces biased results, the question becomes how such situations should be regulated.

Take, for example, the Amazon hiring AI mentioned earlier. Suppose Amazon fed the system with ten years of recruitment data, without any intention of discriminating against female applicants and without manipulating the data. But if the AI simply learned from past recruitment data in which men made up the majority of hires,[14] and thus produced conclusions consistent with those historical ratios, should this be regarded as the AI discriminating against female applicants?

Similar issue can be found in the COMPAS case. If we assume that historical data showed higher crime rates and recidivism among Black individuals than among White individuals,[15] and such data were directly used to train the AI, then the system's assignment of higher recidivism probabilities to Black defendants could be viewed as a rational conclusion based on the available data.[16] The same applies in the economic sphere. In the case of the UK university admissions

---

14 According to reports, 60% of Amazon's workforce and 74% of its management positions were held by men. See James Vincent, "Amazon Reportedly Scraps Internal AI Recruiting Tool That Was Biased against Women," The Verge, October 10, 2018, available at: https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report, cited in Son Young-hwa, "A Study on AI Fairness: Toward a Society with Non-discriminatory AI," Hanyang Law Review, Vol. 34, No. 3, 2023, p. 280.

15 This assumption sets aside the controversies regarding potential flaws in the assumptions or parameters of recidivism prediction algorithms.

16 In reality, some argue that such outcomes are the result of differences in arrest rates ("Why Did Amazon's Hiring AI Favor Men?" Hankook Ilbo, October 14, 2021, available at: https://www.hankookilbo.com/News/Read/A2021101409500001667), while others explain them as stemming from divergent standards of fairness (Haksoo Ko, "The Difficulty of Fair Artificial Intelligence," Bank of Korea Newsletter, August 2021, p. 42 ff.). Whatever the reason, these outcomes consistently reflect underlying social phenomena.

system, there is ample room to interpret the AI's conclusions as rational, since historical data had shown that students from private schools performed better than those from public schools in low-income areas. Of course, the intention here is not to argue that such cases in our society are right, fair, or unbiased. Rather, the question being raised is whether, when focusing on artificial intelligence, the fact that AI produces such outcomes should be treated as a problem of bias warranting its suspension or abandonment. From a technical perspective, one could even argue that AI producing such conclusions reflects a properly designed and adequately trained system that accurately mirrors the intentions of its users or developers.

It is difficult to determine what it truly means for artificial intelligence to be fair[17] or unbiased. It is hard to establish clear standards of fairness. If an outcome turns out to be unfair, does that necessarily make the AI itself unfair? Or, on what grounds and by what standards should fairness or unfairness be judged? These questions remain challenging to answer.[18] In the use of machines or technology, the fairest approach may be the one with the least human involvement. The more humans intervene, the greater the likelihood that their biases—whether intentional or unintentional—will be embedded in the technology. When bias already exists in society, training on such data will inevitably lead to biased outcomes. Because AI is a machine, it can be seen as belonging to the realm of Sein. Likewise, social phenomena and the data derived from them may also be regarded as lying within the realm of Sein rather than Sollen. This raises the question: if AI is fed training data that reflect Sein, is it truly possible to expect it to produce outcomes grounded in Sollen? In other words, can AI be required to adhere to normative demands such as ethics, morality, and law? With current technology, it still seems difficult for AI to recognize and learn Sollen. If so, this issue could in principle be addressed through the selection, manipulation, or processing of the data used to train AI. Yet this immediately raises the question of who should classify and adjust the data, and to what extent. For example, in the Amazon case discussed above, would it be fair to consider training the AI on data in which the ratio of male to female applicants is artificially balanced? Suppose, moreover, that in reality men account for 80–90 percent of perpetrators in sexual crimes. Would an AI trained on data adjusted to a 50:50 ratio between men and women

17 In fact, the concept of fairness is multifaceted and can be interpreted differently across societies, making it difficult to establish a single, definitive definition of what fairness means.

18 Specifically, depending on which standard of fairness is applied to evaluate an algorithm, and on the various contexts in which AI is used, different—and at times even opposing—conclusions about fairness may be drawn. See Haksoo Ko, "The Difficulty of Fair Artificial Intelligence," Bank of Korea Newsletter, August 2021, p. 43.

truly be considered fair? Extending this further, in the case of COMPAS in the United States, if the rate of (re)offending among Black individuals is assumed to be twice as high as that among White individuals, would it be fair to equalize the ratio between Black and White defendants in the training data so that the AI predicts recidivism accordingly? To take an example from our own society rather than from abroad: if news exposure on Naver shows a ratio of conservative to progressive outlets of 7:3, we might call this an unfair service marked by conservative bias. At the very least, there would be grounds to argue that a closer to 5:5 balance between conservative and progressive outlets is needed for fairness. However, if the actual number of conservative and progressive outlets and the ratio of their published articles in society is itself 7:3—given that newspapers are protected in their editorial orientation (Tendenz)—is it fairer to expose users to news at that same 7:3 ratio, or to deliberately adjust it to 5:5? This is a question that is difficult to answer. It is also questionable whether the outcomes produced by AI trained on such arithmetic, quantitative, or outcome-based notions of fairness would possess the level of accuracy necessary for us to accept and employ them, or whether they would even align with the intended purpose of development. Furthermore, one might wonder whether we are placing excessively strong ethical and fairness demands on AI—demands that our own society does not consistently uphold. It is also necessary to consider how to address the new forms of bias that may arise from the selection and processing of training data.

C. Algorithmic Bias

One possible approach to address concerns on manipulating training data is to explore ways of designing algorithms that explicitly take such issues into account. While training data may belong more to the realm of Sein, the design of algorithms is carried out by humans and therefore lies within a domain where Sollen can be introduced. Accordingly, it is possible to mandate or require that such considerations be incorporated into the design process. However, what is troubling is that bias arising from algorithms can lead to even more dangerous outcomes, since it involves artificial adjustments of results through human intervention.

Algorithmic bias typically arises from three main sources. The first is when the algorithm is intentionally designed to be biased at the design stage. The second is when, although not intentional, the developer's unconscious biases are nonetheless reflected in the design.[19] The

19 Lee, Joon-Il, Artificial Intelligence and the Constitution, Constitutional Studies Vol. 28, No. 2, 2022, p. 365;Park, Do-Hyun, The Intersection of Human Bias and Artificial Intelligence, Seoul National University Law Review Vol. 63, No. 1, 2022, p. 151 (distinguishing between "conscious" and "implicit" biases).

third, which may be difficult to distinguish from the second, occurs when social discrimination issues are not taken seriously in the design process—that is, when existing biases are left uncorrected.[20] (This, in turn, can interact with the training data and manifest as bias in the outcomes.)

When discrimination arises because algorithms are deliberately designed to be biased, the consequences are even more serious than those stemming from biased training data. If AI merely mirrors the biases already present in society, such outcomes may be tolerated to some extent. However, discrimination caused by algorithmic bias constitutes an artificial form of discrimination. For example, in the financial sector, there is a fundamental difference between developing AI for loan assessments (such as interest rate calculations) or credit scoring based on existing risk-factor data, and designing algorithms to weight variables such as being female or being Black. Such practices could likely be fall under existing laws that regulate or prohibit discriminatory conduct.

The second case poses the problem that such biases are difficult to detect. This stems from the invisibility[21] and opacity of the computational processes that drive algorithms, making them hard to explain. The opacity of algorithms may result from developers' reluctance to disclose trade secrets, from the lack of expertise among non-specialists to understand how they operate, or even from the difficulty that experts themselves face in comprehending their workings.[22] For instance, Naver had already been criticized for bias in the placement of political news. In response to such public criticism, Naver announced in 2018 that news editing would no longer be performed by humans but would instead be assigned to AI, with humans merely providing technical support. Nevertheless, because many continue to criticize Naver's news placement as showing a conservative bias, suspicions persist about its news curation

20 Kim, Sung-Yong & Jung, Kwan-Young, A Study on Discrimination Caused by Automated Processing of Personal Data by Artificial Intelligence, Seoul National University Law Review, Vol. 60, No. 2, 2019, p. 326.

21 Jeong, Won-Seob, Bias and Fairness in Artificial Intelligence Algorithms, Human · Environment · Future, No. 25, 2020, p. 65.

22 Won, Sang-Cheol, The Intersection of Artificial Intelligence Ethics and Law, Journal of Legal Theory and Practice Vol. 12, No. 2, 2024, p. 213;Kim, Ji-Yeon, The Ethical Status of Artificial Intelligence (AI): The Ethical Niche of Artificial Intelligence: Mingling with Humans and Non-Humans, Society and Theory No. 46, 2023.

algorithms. However, Naver has refused to disclose the technology, citing trade secrets.[23]

The third case raises the question of whether algorithms can be normatively required to be designed or modified in ways that eliminate or reduce biases arising from training data. While algorithmic adjustments may help address bias in AI, they also risk creating new equality concerns stemming from algorithmic bias itself. Regulating the design or manipulation of algorithms is by no means an easy task. Since the algorithms used to design and train AI are developed with substantial investment of time, resources, and efforts of developers or companies, they are generally treated as trade secrets, and disclosure is highly likely to be refused. It is also extremely difficult to mandate disclosure or to require developers to explain or reveal the development process. This is often framed in terms of the duty to explain, explainability, the right to explanation, or algorithmic transparency. However, in practice, these principles appear very difficult to implement.

For example, in 2022 the Seoul High Court ruled against Naver in a case where the company was accused of manipulating its algorithms to favor its own services, leading the Korea Fair Trade Commission (KFTC) to impose corrective measures and a fine.[24] Naver, however, appealed to the Supreme Court, arguing that modifying algorithms is a routine practice in search engines and that such adjustments were intended to enhance consumer utility.[25] Likewise, Coupang was fined 140 billion KRW by the KFTC for allegedly manipulating its algorithms to unfairly favor its own private brand (PB) products, but the company has argued that the algorithmic adjustments were legitimate business practices.[26]

These examples highlight both the risks of algorithmic manipulation and the challenges of regulating it. The cases above were not merely about the fairness of the algorithms themselves

---

23 "Revealed Conservative Bias in Naver's News Editing AI Algorithm… Will the Fairness Debate Heat Up?" AI Times, March 8, 2021, https://www.aitimes.com/news/articleView.html?idxno=137148.

24 Seoul High Court, Decision 2021Nu36129, December 14, 2022.

25 "'Comparison Shopping Search Algorithm Manipulation Charges'—FTC's KRW 26.6 Billion Fine on Naver Upheld," Law Times, December 15, 2022, https://www.lawtimes.co.kr/news/183827;"Naver and the FTC in Legal Dispute over the 'Fairness of Search Algorithms,'" Maeil Ilbo, February 19, 2023, https://www.m-i.kr/news/articleView.html?idxno=989013.

26 "Coupang Manipulated Algorithms to Rank Its Own Products First… Employees Wrote 70,000 Reviews," Dong-A Ilbo, June 13, 2024, https://www.donga.com/news/Economy/article/all/20240613/125411557/2.

but involved deliberate interventions by corporate insiders for economic gain. Nonetheless, the companies argued that their actions were normal and unavoidable measures taken to improve performance. Moreover, because companies are reluctant to disclose their algorithms on grounds such as copyright and trade secrets, it is difficult to determine how such outcomes were produced. These cases came to light only because relevant laws, investigations, and evidence were in place. By contrast, in situations involving algorithmic manipulation leading to discrimination on the basis of gender, race, or other characteristics, detecting such biases—and proving them—becomes far more difficult.

How can proactive intervention—deliberately curating and manipulating training data and designing or adjusting algorithms to correct and mitigate existing biases—be justified in the name of AI fairness? Moreover, if such measures are to be adopted, we must also debate what level of equality is required for an AI system to be regarded as fair.

To begin with, this debate is normatively analogous to discussions surrounding affirmative action. Affirmative action refers to temporary preferential measures taken by States to favor groups that have historically suffered discrimination, in order to compensate for past disadvantages. A common example is gender quotas.[27] Similarly, although inequalities persist in society, one could argue that when feeding training data into AI systems, priority should be given to historically marginalized groups. The idea is not to design value-neutral algorithms, but rather to develop value-oriented algorithms that give precedence to, or at least ensure equal treatment with, those who have been subject to discrimination in society.

And there must also be discussion about whether fairness measures in AI should be implemented on the basis of an arithmetic average. For example, even if crime rates are statistically higher among certain groups—such as Black or White individuals, or men or women—should training data nevertheless be balanced to reflect equal proportions across race and gender? Or should algorithms be adjusted so that their outcomes produce such equal proportions? The same questions arise in employment and in political orientation. As noted earlier, in the case of news portals, the issue would be whether news exposure should be presented at a 50:50 ratio between progressive and conservative outlets. If, despite the actual distribution of public opinion, an arithmetic balance is imposed on what is shown, this could

27 However, such affirmative measures may in turn generate inequalities for other groups. It must therefore be carefully reconsidered to what extent these measures can be justified and to what degree they are constitutionally permissible.

in fact be seen as AI distorting or manipulating human decision-making rather than fostering fairness.[28]

Ⅳ. Measures to Ensure Fairness in Artificial Intelligence

The best way to address the problem of bias in artificial intelligence, as explained earlier, is to eliminate bias in our society itself. The next best way is for algorithm developers to be free from bias or to actively seek to correct it. In the end, discussions on AI ethics so far have been less about instilling ethics or sollen in AI machines themselves, and more about demanding ethical responsibility from developers.

In order to ensure fairness in artificial intelligence, various regulatory approaches have been proposed at different stages, ranging from the input of training data to the design of algorithms. Examples include discussions on data governance related to training data, as well as obligations of explanation and rights to request explanation concerning algorithms. The following section explores measures taken by various countries to secure fairness in AI.

A. EU

The European Union enacted the Artificial Intelligence Act to regulate discrimination in AI. To this end, it emphasizes transparency and prohibits the use of biometric categorization systems that classify natural persons individually on the basis of biometric data to infer race, political opinions, trade union membership, religious or philosophical beliefs, sexual life, or sexual orientation. It also prohibits AI systems that exploit the vulnerabilities of individuals or groups due to their age, disability, or specific social or economic situation, as well as AI systems that evaluate or classify individuals or groups over a certain period of time based on their social behavior, or on known, inferred, or predicted personality traits or characteristics, through social scoring. (Article 5)

Moreover, in order to regulate such high-risk AI systems, requirements are imposed on data collection and labelling, the availability, quantity, and suitability of the necessary data,

28 Muibin, A Re-examination of Freedom of Thought: Focusing on the Possibility of Manipulating the Decision-Making Process by Artificial Intelligence, Public Law Review, Vol. 52, No. 3, 2024.

as well as assessments thereof. It also mandates appropriate measures to identify, prevent, and mitigate potential biases that may lead to discrimination prohibited under EU law. In particular, Article 10 establishes provisions on data and data governance, aiming to regulate quality standards for training data.

B. USA

According to the 2022 National Artificial Intelligence Initiative Act of the United States, the National Artificial Intelligence Advisory Committee was established to address these issues. In October 2022, the "AI Bill of Rights"[29] was published, and on October 30, 2023, the Biden Administration issued the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[3031]. Recognizing that AI carries inherent risks along with its many potential benefits, this Executive Order set out eight guiding principles[32] that federal agencies must follow in the development and use of AI. It also mandated the establishment of a White House AI Council[33] to formulate and implement related policies and to coordinate government activities.

At the private sector level, the Algorithmic Accountability Act of 2022 seeks to prevent and regulate problems arising from errors or biases in the use of algorithms. To this end, the Act introduces safeguards in the form of mandatory impact assessments and empowers the

---

29 https://www.whitehouse.gov/ostp/ai-bill-of-rights/

30 Executive Order 14110 of October 30, 2023 — Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

31 https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

32 Executive Order 14110 of October 30, 2023 sets out eight guiding principles: Ensuring the Safety and Security of AI Technology, Promoting Innovation and Competition, Supporting Workers, Advancing Equity and Civil Rights, Protecting Consumers, Patients, Passengers, and Students, Protecting Privacy, Advancing Federal Government Use of AI, and Strengthening American Leadership Abroad.

33 Ministry of Government Legislation, Future Legislative Innovation Planning Team, Domestic and International Legislative Trends on Artificial Intelligence (AI), Legislation Newsletter, 2024, p. 31.

Federal Trade Commission to ensure their effectiveness.[34] The United States has adopted a dual approach: on one hand, creating advisory bodies to flexibly respond to emerging issues, and on the other, enforcing binding legal frameworks that apply separately to the public and private sectors.

C. OECD

The OECD issued the Recommendation of the Council on Artificial Intelligence in 2019, and released an updated version in May 2024. The Recommendation emphasizes that AI systems must not lead to unjust discrimination in areas such as employment, finance, and healthcare. It points out that AI can produce unfair outcomes not only through intentional bias but also through unintentional bias, and recommends that AI developers conduct prior analyses of risks of bias and unfairness and establish corrective measures. The Recommendation also stresses transparency and explainability, urging AI developers and operators to practice responsible disclosure and to ensure that those affected by AI decisions have the right to challenge outcomes or request explanations.

Ⅴ. Conclusion

Around the world, there is growing recognition of the adverse effects and inequalities caused by bias in artificial intelligence, and efforts are being made to address these issues through both hard and soft forms of regulation.

However, as examined above, while some types of bias can be addressed within the domain of regulation, others remain hidden, reproducing existing inequalities in society or even exacerbating them.

Biases that have already been widely discussed and recognized as problems seem to have found at least some paths toward resolution. For instance, cases where data lacks representativeness or where training data has been incorrectly input from the outset are, in fact, relatively easier to remedy.

---

34 For more detailed provisions on equal protection under the Algorithmic Accountability Act, see Ilwoo Kim, A Study on Discrimination in High-Risk AI Systems, Sogang Law Review, Vol. 13, No. 1, 2024, p. 33.

However, there appears to be insufficient discussion on how to address situations where there are no problems in the training data or in the algorithm itself, yet the outcomes of artificial intelligence still turn out to be biased and unequal. A representative example is when AI learns the existing biases of our society—such as discrimination against women or racial minorities—and produces biased conclusions accordingly. The unresolved question is whether such AI systems should be considered unfair.

In practice, even when neither the training data nor the algorithm contains flaws, there may be cases where AI produces biased or unequal outcomes. This raises potential legal questions, such as whether the use of such AI can be forcibly prohibited, and how compensation or recovery of development costs should be addressed.

Can we command AI to learn sollen so that it autonomously corrects the biases inherent in our society and produces ethically and normatively fair conclusions? Is it even possible for AI to learn sollen? How can we determine which conclusions are fair and unbiased? To what extent must training data or algorithms be adjusted before an AI system can be regarded as unbiased?

As of yet, it seems difficult for our society to provide a definitive or common answer to these questions. We still need more extensive discussion, reflection, and social consensus on issues of fairness, bias, and inequality in artificial intelligence.

It is crucial, however, that before AI becomes more widely commercialized, we must engage in this reflection and establish at least some form of guidelines—whether through normative (mandatory) regulation or through self-regulatory mechanisms. Otherwise, once AI begins to learn the biases of our society without any understanding of sollen, it will inevitably reproduce and even amplify existing inequalities, risking the nullification of the progress our society has thus far made toward reducing inequality.

[발표 2 | Speaker 2]

데이터, 사생활 보호, 그리고 차별

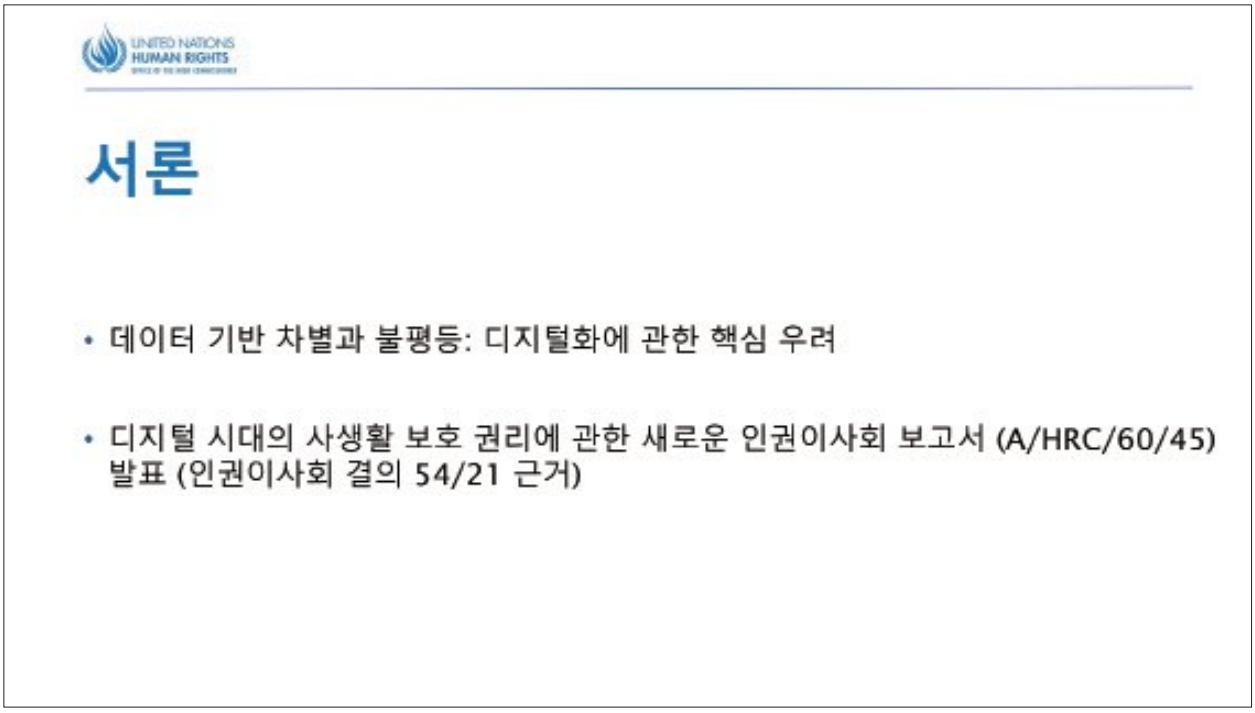
Data, Privacy and Discrimination

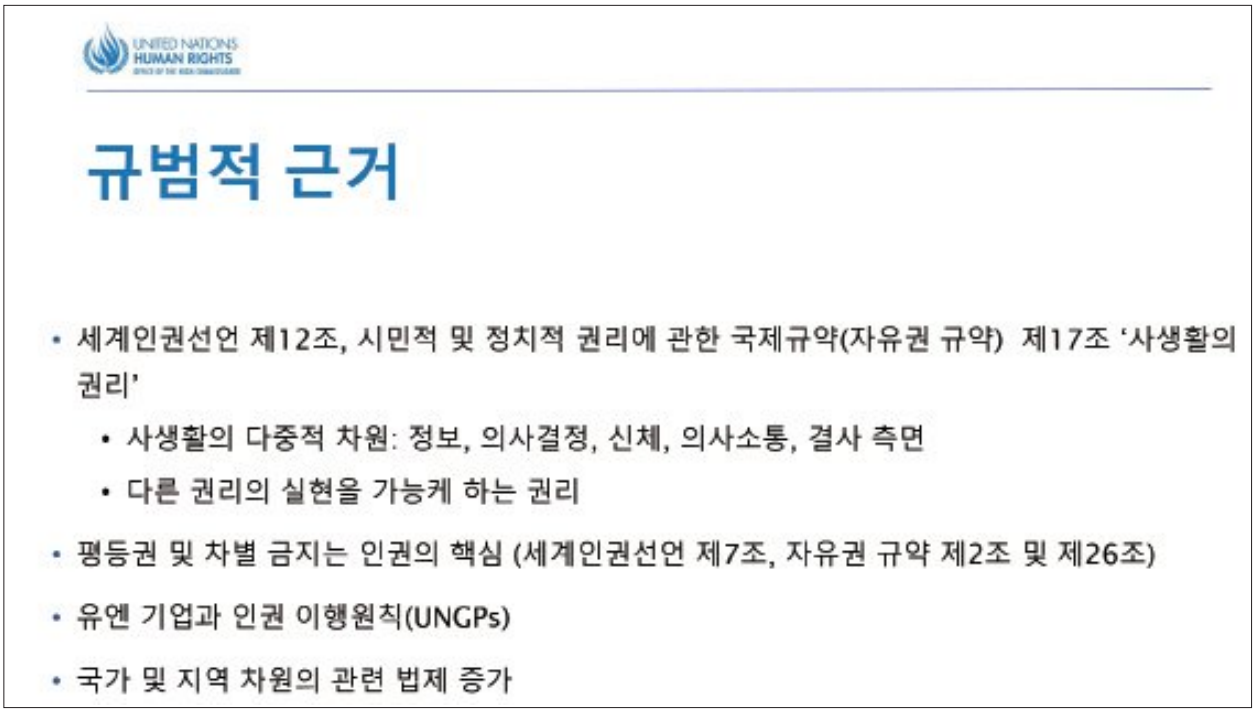
팀 엥겔하르트
Tim ENGELHARDT

유엔 인권최고대표사무소 인권담당관
Human Rights Officer, OHCHR

데이터, 사생활 보호, 그리고 차별

팀 엥겔하르트 | 유엔 인권최고대표사무소 인권담당관


UNITED NATIONS
HUMAN RIGHTS
서론
• 데이터 기반 차별과 불평등: 디지털화에 관한 핵심 우려
• 디지털 시대의 사생활 보호 권리에 관한 새로운 인권이사회 보고서 (A/HRC/60/45) 발표 (인권이사회 결의 54/21 근거)


UNITED NATIONS
HUMAN RIGHTS
규범적 근거
• 세계인권선언 제12조, 시민적 및 정치적 권리에 관한 국제규약(자유권 규약) 제17조 '사생활의 권리'
• 사생활의 다중적 차원: 정보, 의사결정, 신체, 의사소통, 결사 측면
• 다른 권리의 실현을 가능케 하는 권리
• 평등권 및 차별 금지는 인권의 핵심 (세계인권선언 제7조, 자유권 규약 제2조 및 제26조)
• 유엔 기업과 인권 이행원칙(UNGPs)
• 국가 및 지역 차원의 관련 법제 증가


UNITED NATIONS
HUMAN RIGHTS
핵심 영역에서의 차별과 불평등
• 법 집행 및 형사 사법
• 이주
• 사회 복지
• 보건
• 디지털 공공 인프라
• 감시
• 온라인 정보


UNITED NATIONS
HUMAN RIGHTS
도전 과제: 시스템 설계 상의 문제
• 데이터 편향과 설계상 선택의 편향
• 투명성과 책임성 부족
• 확장성
• 추론 및 예측

UNITED NATIONS HUMAN RIGHTS

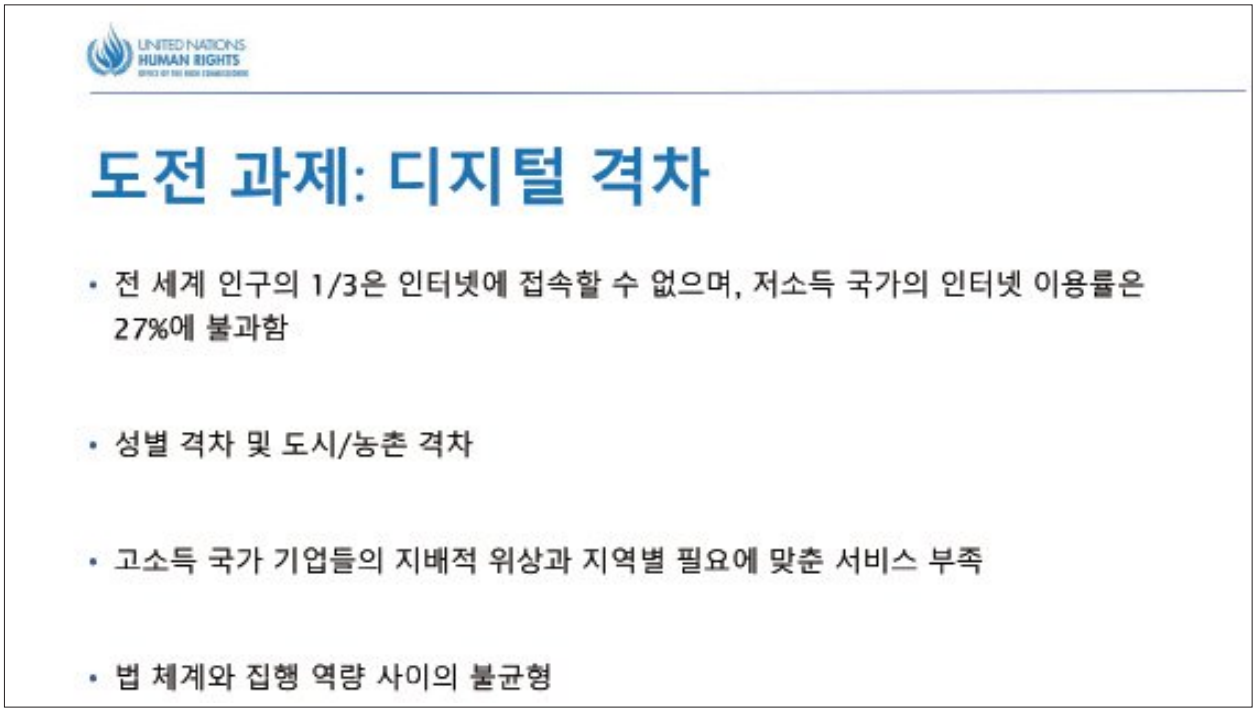
도전 과제: 인간, 구조, 제도적 편향

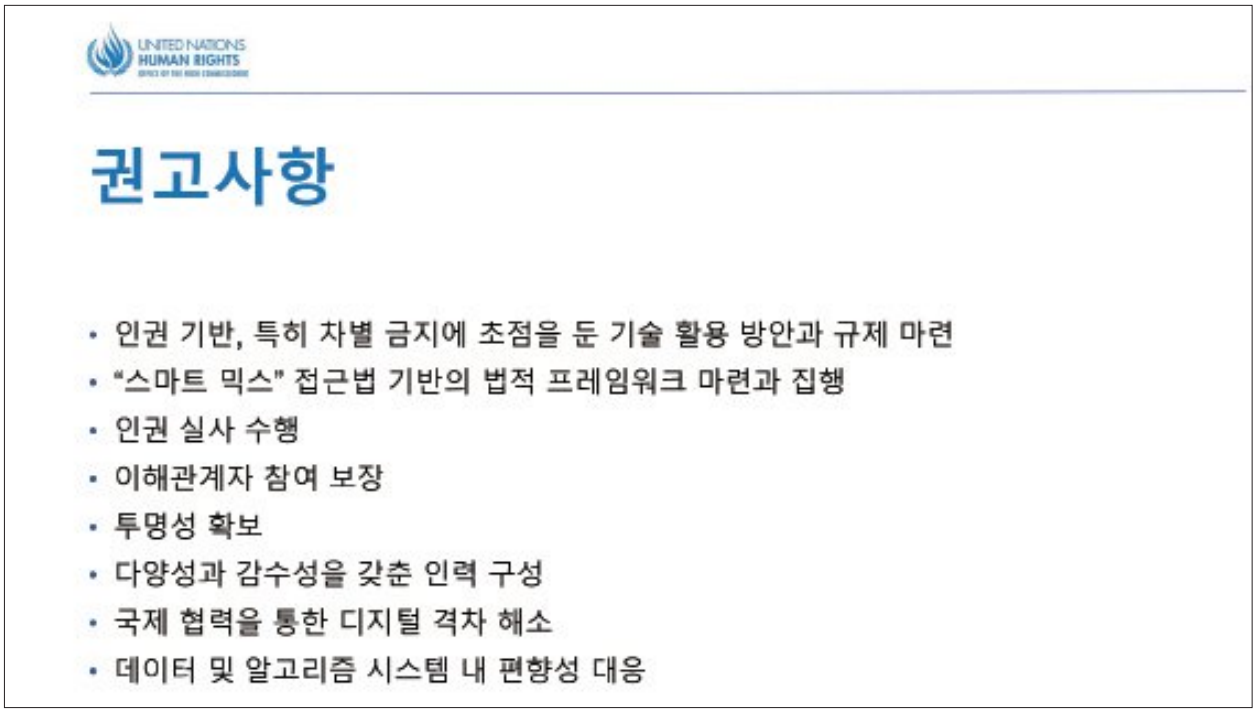
- 오랜 시간 이어진 불의, 인간의 뿌리 깊은 편향과 추정이 반영된 시스템의 편향성
- 교차성
- 데이터가 중립적이라는 인식이 실재하는 불평등을 가리는 문제

UNITED NATIONS HUMAN RIGHTS

도전 과제: 민간 부문의 역할

- 기업이 기술 발전을 주도
- 국가의 보호 의무와 기업의 존중 책임 (유엔 기업과 인권 이행원칙, UNGP)
- 국가와 기업의 밀접한 연계가 초래하는 과제


UNITED NATIONS
HUMAN RIGHTS
도전 과제: 디지털 격차
• 전 세계 인구의 1/3은 인터넷에 접속할 수 없으며, 저소득 국가의 인터넷 이용률은 27%에 불과함
• 성별 격차 및 도시/농촌 격차
• 고소득 국가 기업들의 지배적 위상과 지역별 필요에 맞춘 서비스 부족
• 법 체계와 집행 역량 사이의 불균형


UNITED NATIONS
HUMAN RIGHTS
권고사항
• 인권 기반, 특히 차별 금지에 초점을 둔 기술 활용 방안과 규제 마련
• "스마트 믹스" 접근법 기반의 법적 프레임워크 마련과 집행
• 인권 실사 수행
• 이해관계자 참여 보장
• 투명성 확보
• 다양성과 감수성을 갖춘 인력 구성
• 국제 협력을 통한 디지털 격차 해소
• 데이터 및 알고리즘 시스템 내 편향성 대응


UNITED NATIONS
HUMAN RIGHTS
감사합니다
문의:
tim.engelhardt@un.org

Data, Privacy and Discrimination

Tim ENGELHARDT | Human Rights Officer, OHCHR

UNITED NATIONS HUMAN RIGHTS

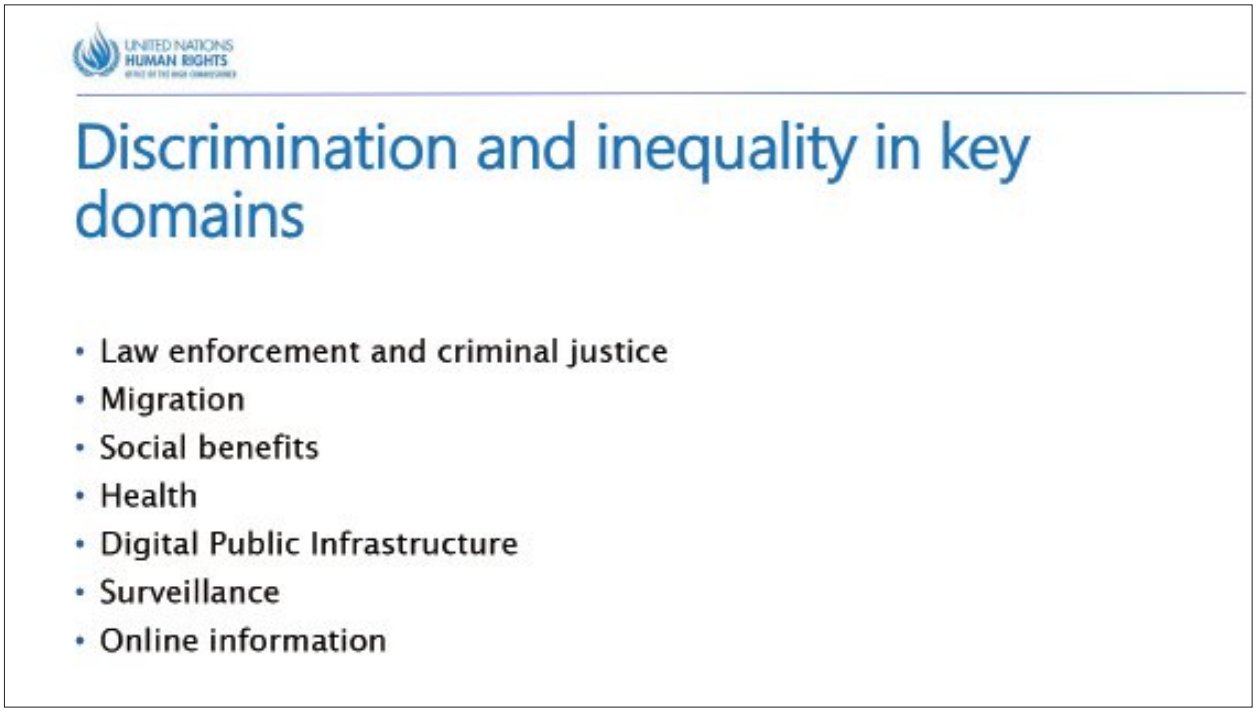
Introduction

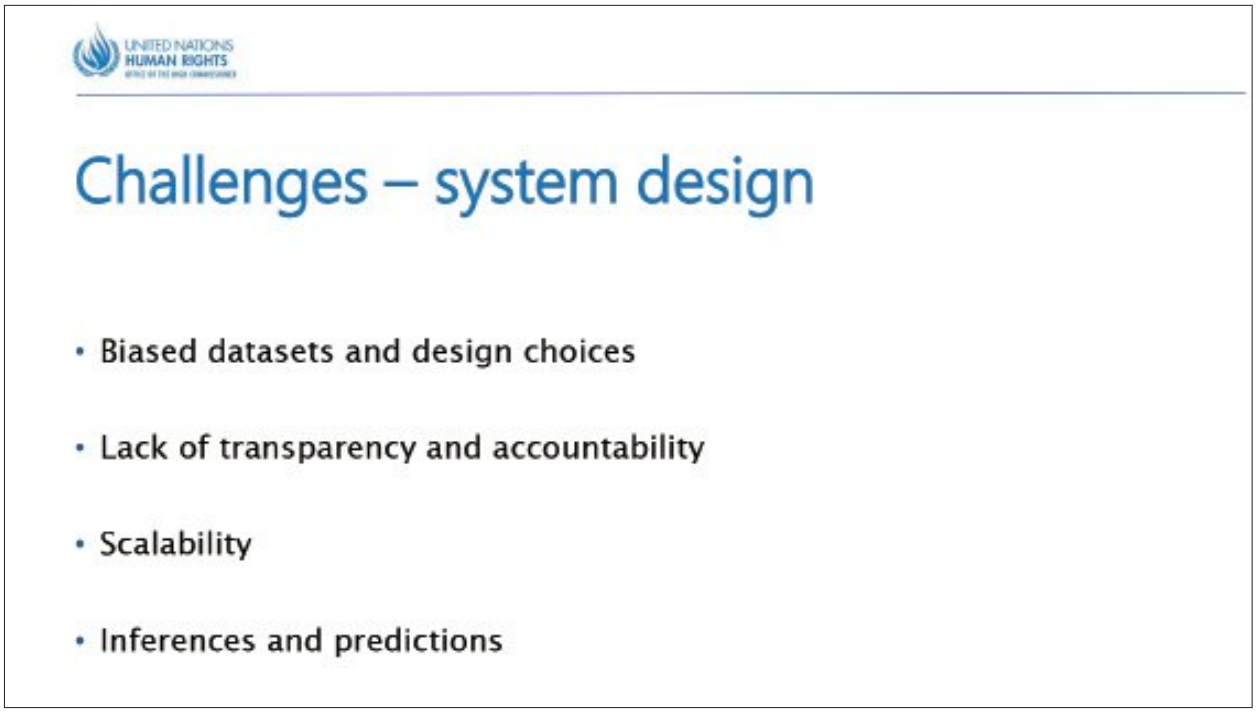
- Data-driven discrimination and inequality a key concern linked to digitalization
- New report on the right to privacy in the digital age A/HRC/60/45 (based on HRC resolution 54/21)

UNITED NATIONS HUMAN RIGHTS

Normative Framework

- Right to privacy, UDHR, article 12, ICCPR, article 17
  - Multiple privacy dimensions: informational, decisional, bodily, communication, associational
  - Enabler of other rights
- Right to equality and prohibition of discrimination core to human rights (UDHR, article 7, ICCPR; articles 2 and 26)
- UN Guiding Principles on Business and Human Rights
- Growing number of national and regional legal frameworks


UNITED NATIONS
HUMAN RIGHTS
Discrimination and inequality in key domains
• Law enforcement and criminal justice
• Migration
• Social benefits
• Health
• Digital Public Infrastructure
• Surveillance
• Online information


UNITED NATIONS
HUMAN RIGHTS
Challenges – system design
• Biased datasets and design choices
• Lack of transparency and accountability
• Scalability
• Inferences and predictions

UNITED NATIONS HUMAN RIGHTS

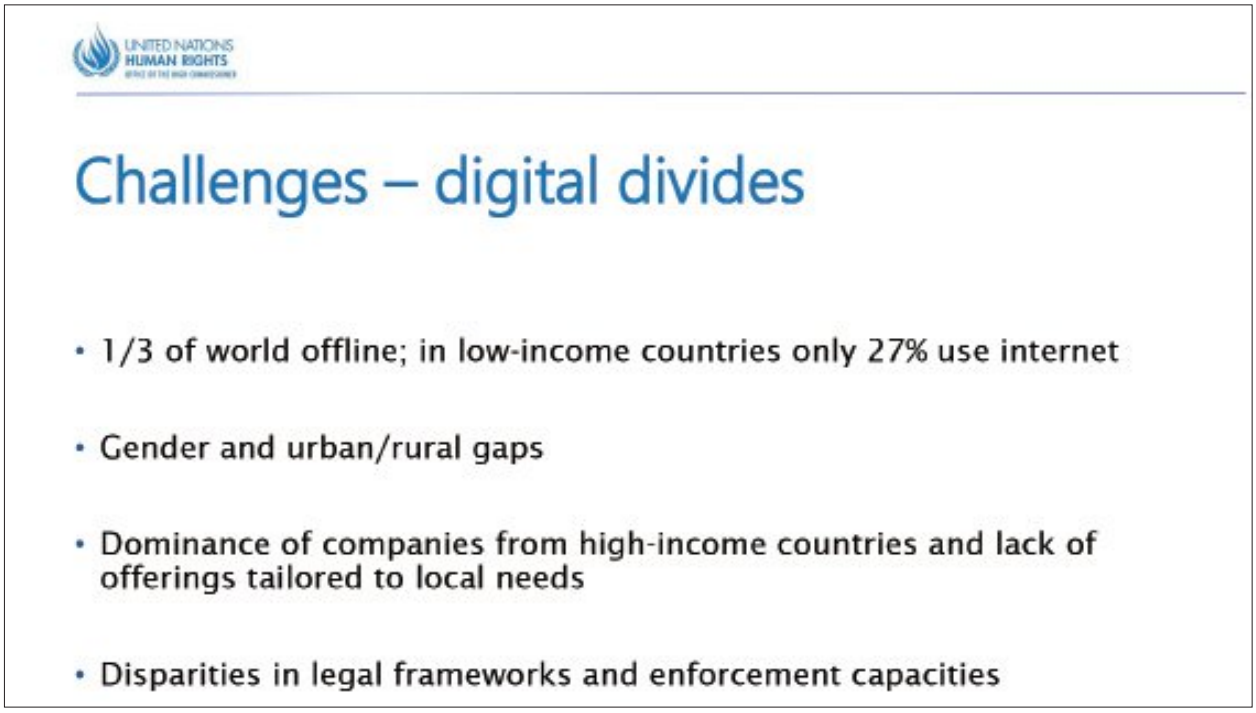
Challenges – human, structural and institutional bias

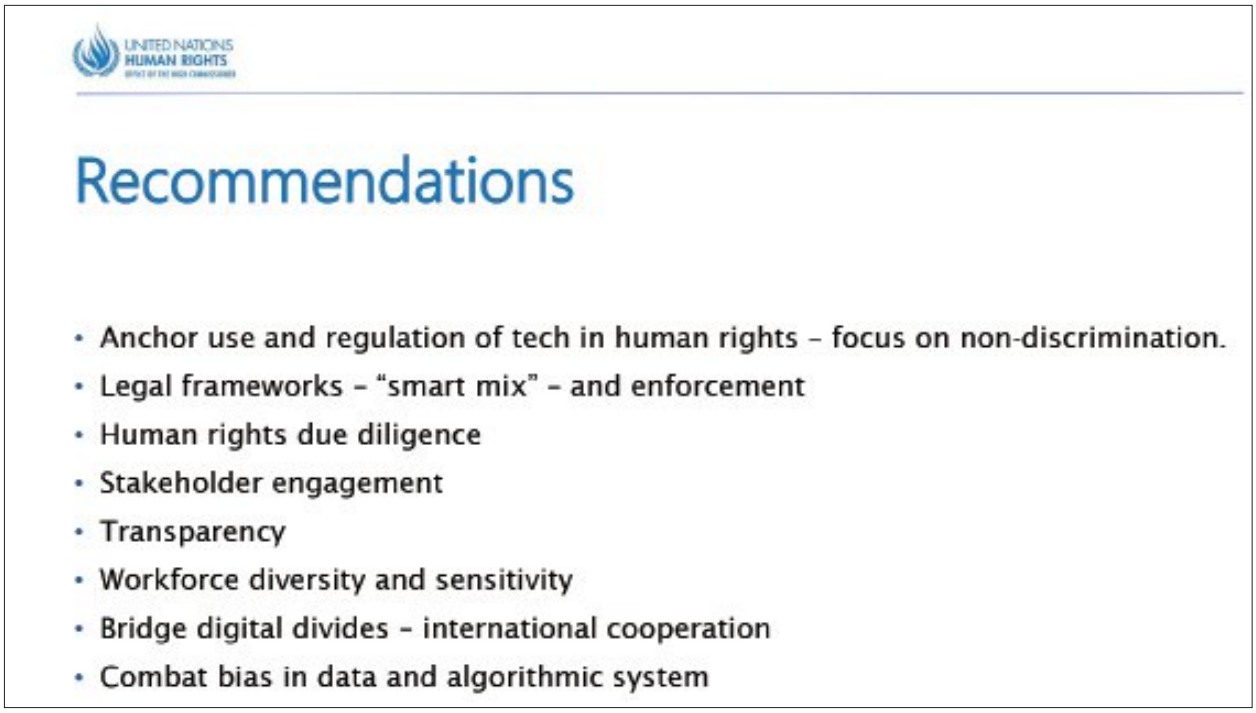
- Systemic bias reflecting historic injustices, deep-seated human bias and assumptions
- Intersectionality
- Perceived objectivity of data obfuscating unfairness of practices

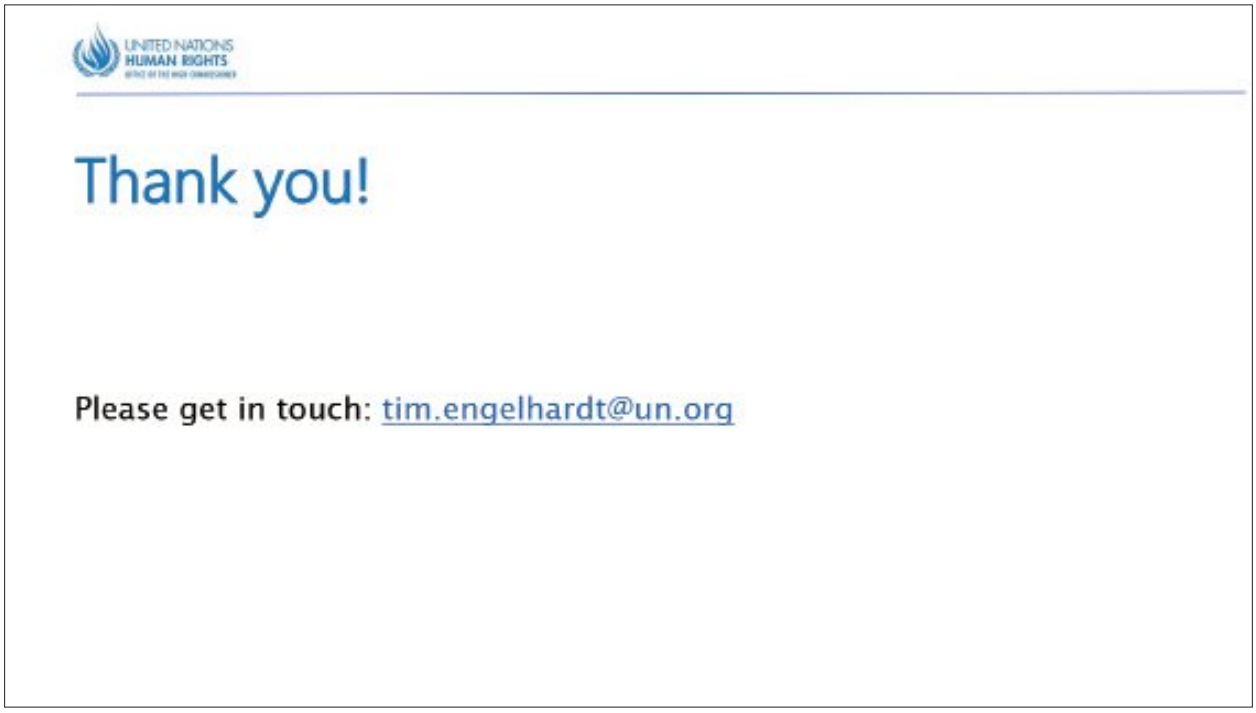
UNITED NATIONS HUMAN RIGHTS

Challenges - role of the private sector

- Businesses drive tech development
- State duty to protect, business responsibility to respect (UN Guiding Principles on Business and Human Rights)
- Close State-business nexus creates particular challenges


UNITED NATIONS
HUMAN RIGHTS
Challenges – digital divides
• 1/3 of world offline; in low-income countries only 27% use internet
• Gender and urban/rural gaps
• Dominance of companies from high-income countries and lack of offerings tailored to local needs
• Disparities in legal frameworks and enforcement capacities


UNITED NATIONS
HUMAN RIGHTS
Recommendations
• Anchor use and regulation of tech in human rights – focus on non-discrimination.
• Legal frameworks – "smart mix" – and enforcement
• Human rights due diligence
• Stakeholder engagement
• Transparency
• Workforce diversity and sensitivity
• Bridge digital divides – international cooperation
• Combat bias in data and algorithmic system


UNITED NATIONS
HUMAN RIGHTS
Thank you!
Please get in touch: tim.engelhardt@un.org

[발표 3 | Speaker 3]

기술 발전과 디지털 격차

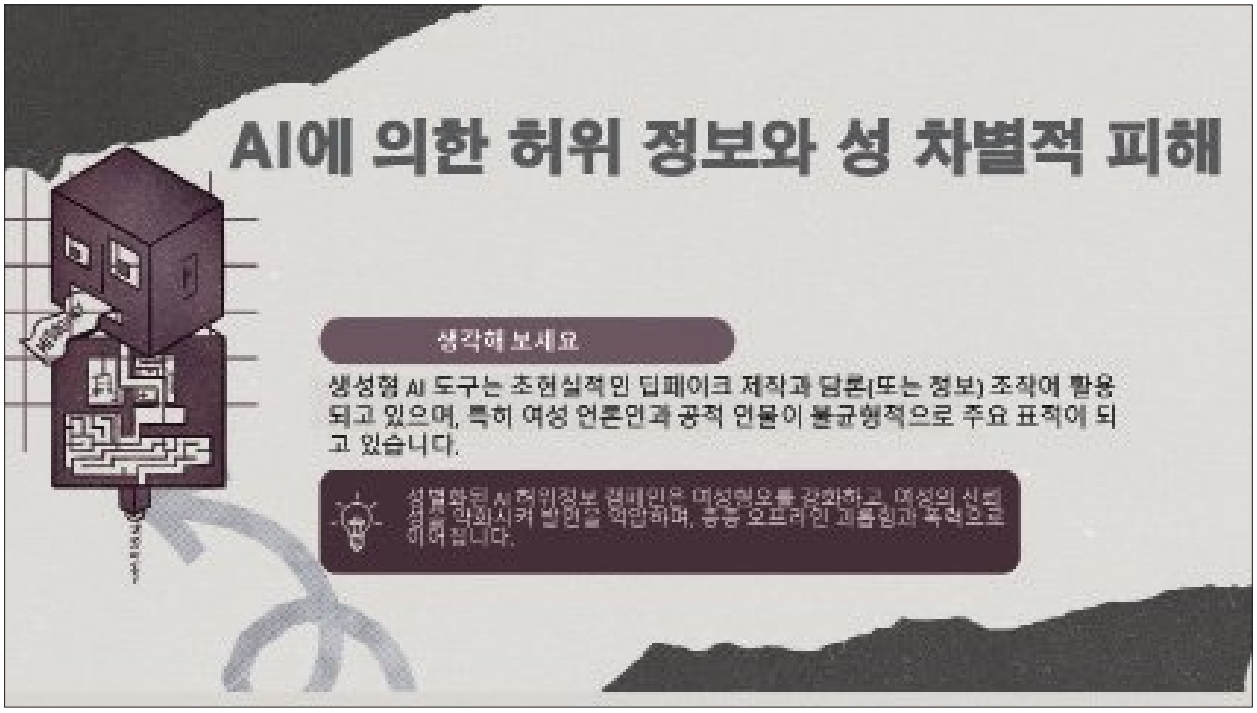
Technological Development and the Digital Divide

나이갓 다드
Nighat DAD

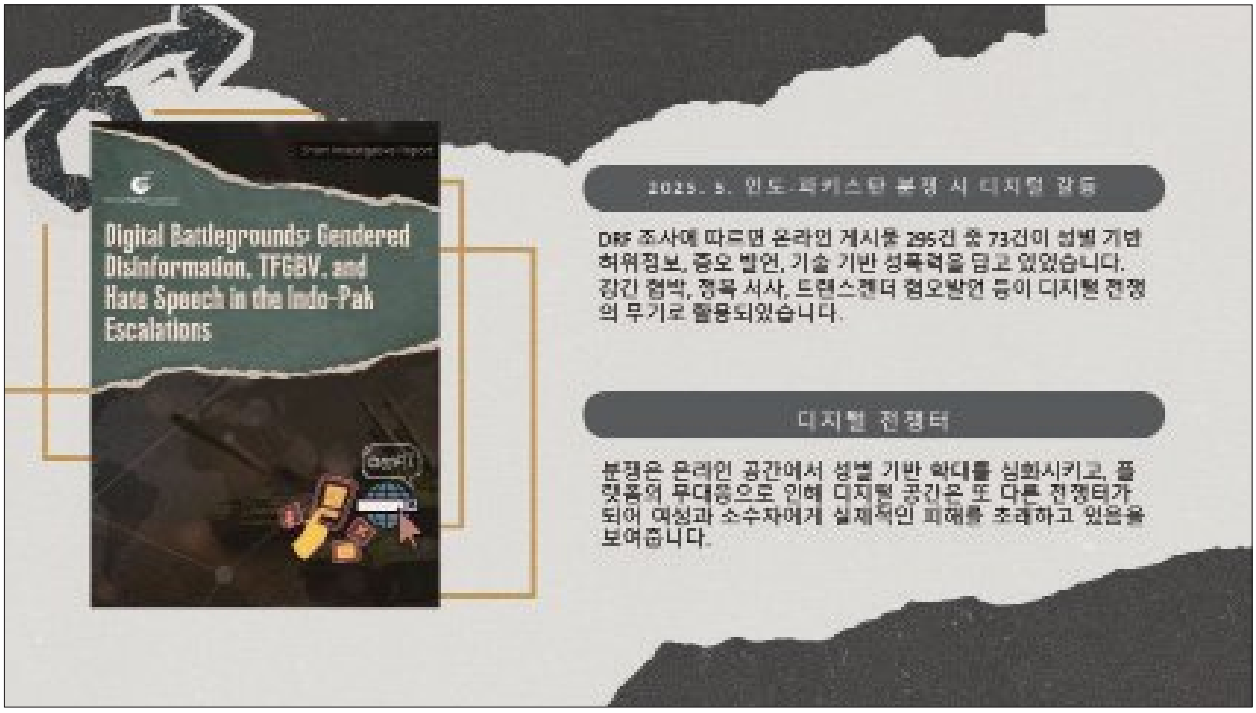
디지털 권리 재단 상임이사, 유엔 AI 자문기구 위원
Executive Director, Digital Rights Foundation
Member, UN High-level Advisory Body on AI

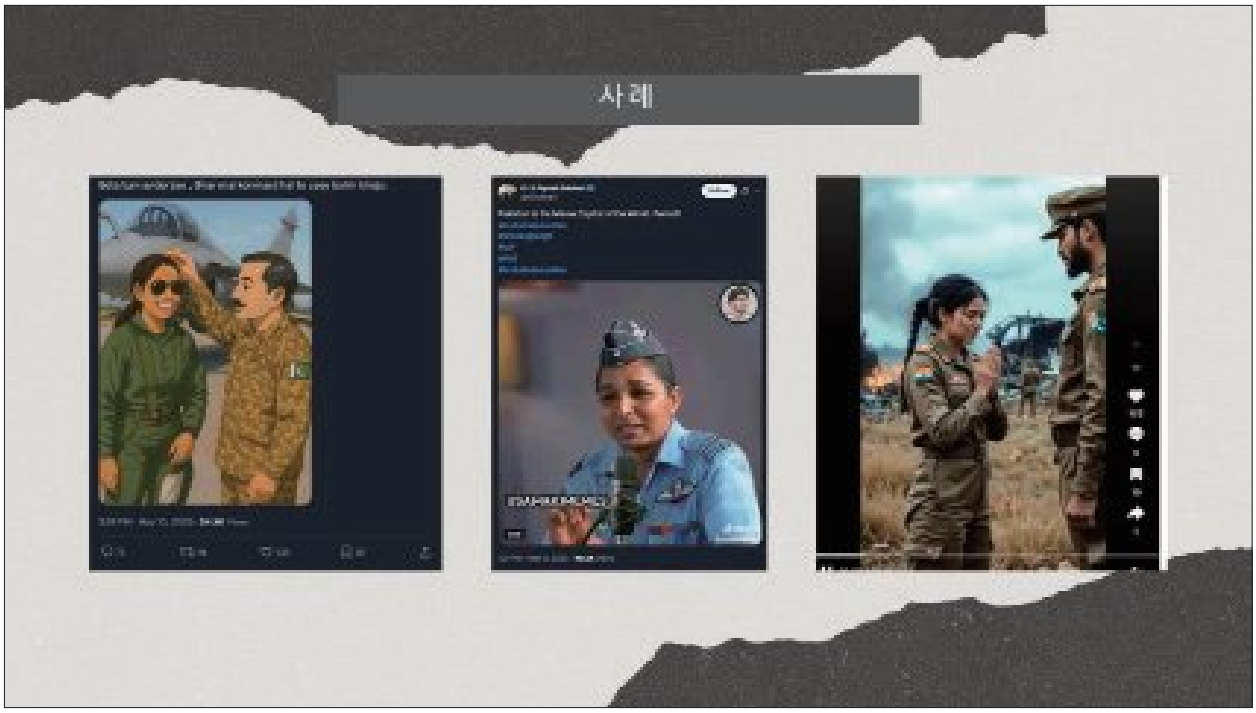
기술 발전과 디지털 격차

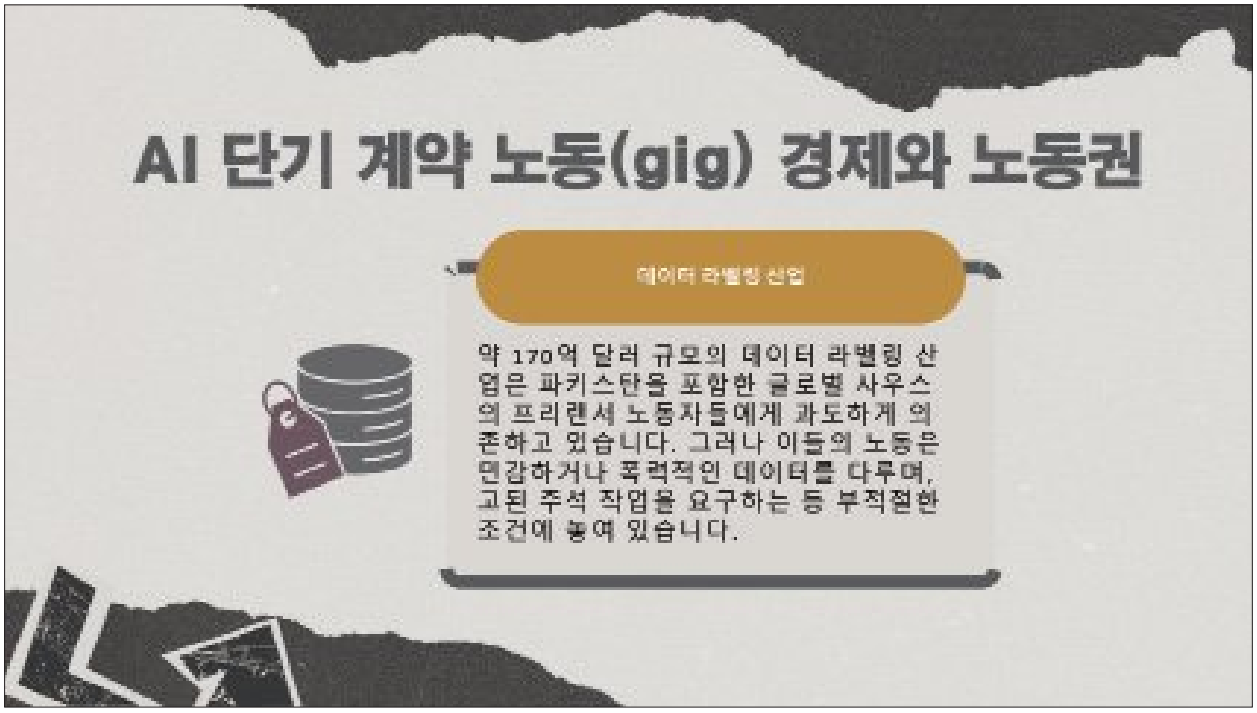
나이갓 다드 | 디지털 권리 재단 상임이사, 유엔 AI 자문기구 위원

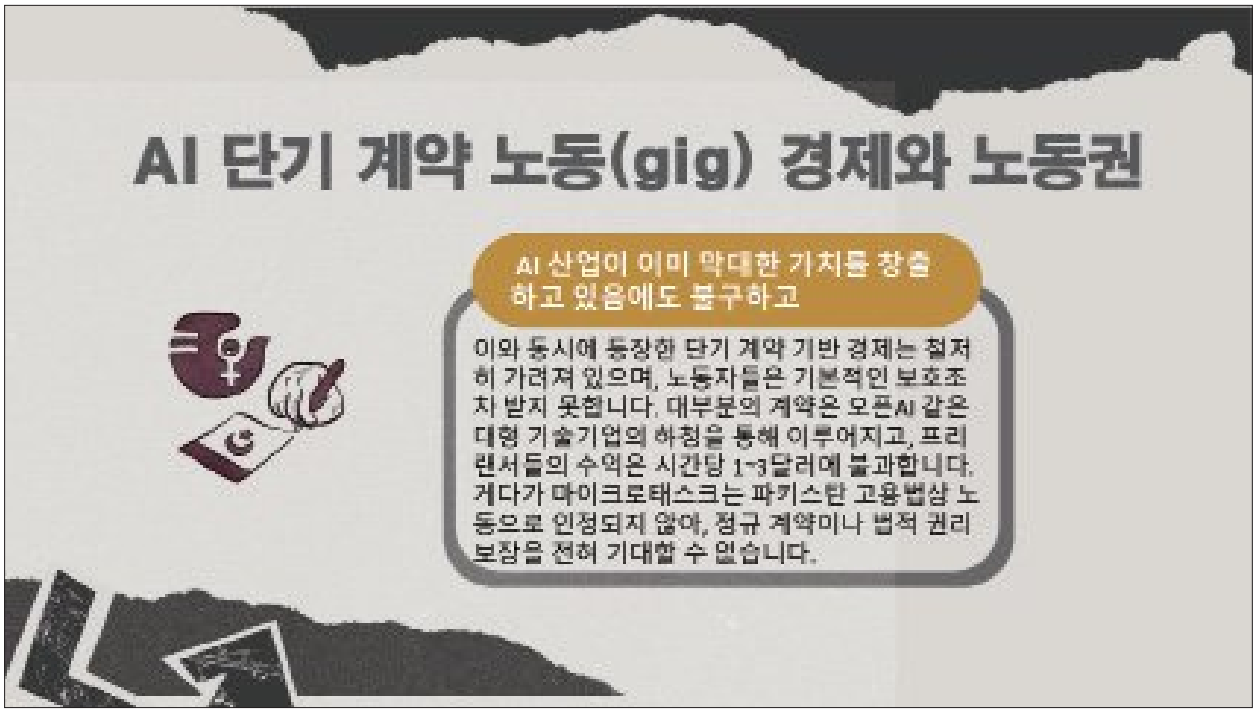

AI에 의한 허위 정보와 성 차별적 피해
생각해 보세요
생성형 AI 도구는 초현실적인 딥페이크 제작과 담론(또는 정보) 조작에 활용되고 있으며, 특히 여성 언론인과 공적 인물이 불균형적으로 주요 표적이 되고 있습니다.
성별화된 AI 허위정보 캠페인은 여성혐오를 강화하고, 여성의 신뢰성을 약화시키며 실현을 억압하며, 궁극적으로 온라인 괴롭힘과 폭력으로 이어집니다.

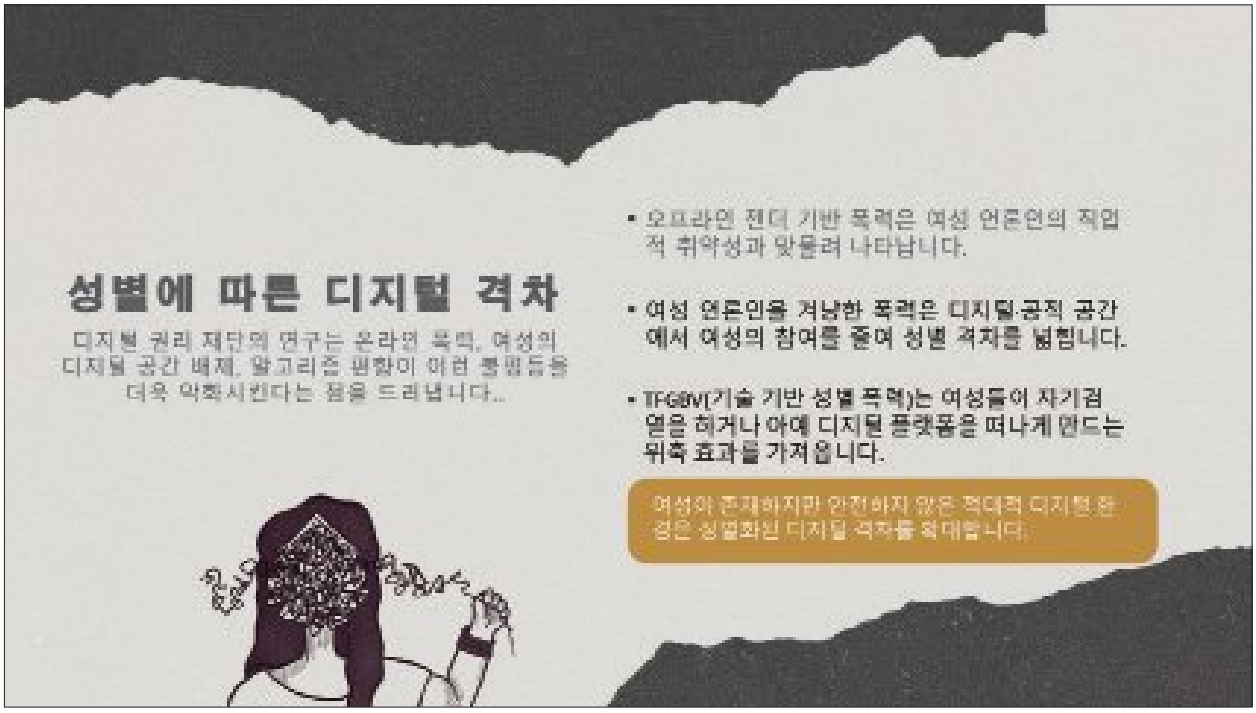

명예 살인, 인공지능이 조장하는 성별 기반 폭력
NEW TURN IN BALOCHISTAN CASE
AI 시스템의 안전장치와 투명성이 부족해, 미디 접근성과 디지털 문해력에서 장벽을 겪는 파키스탄 여성들은 새로운 형태의 기술 기반 학대에 더 쉽게 노출되고 있습니다.

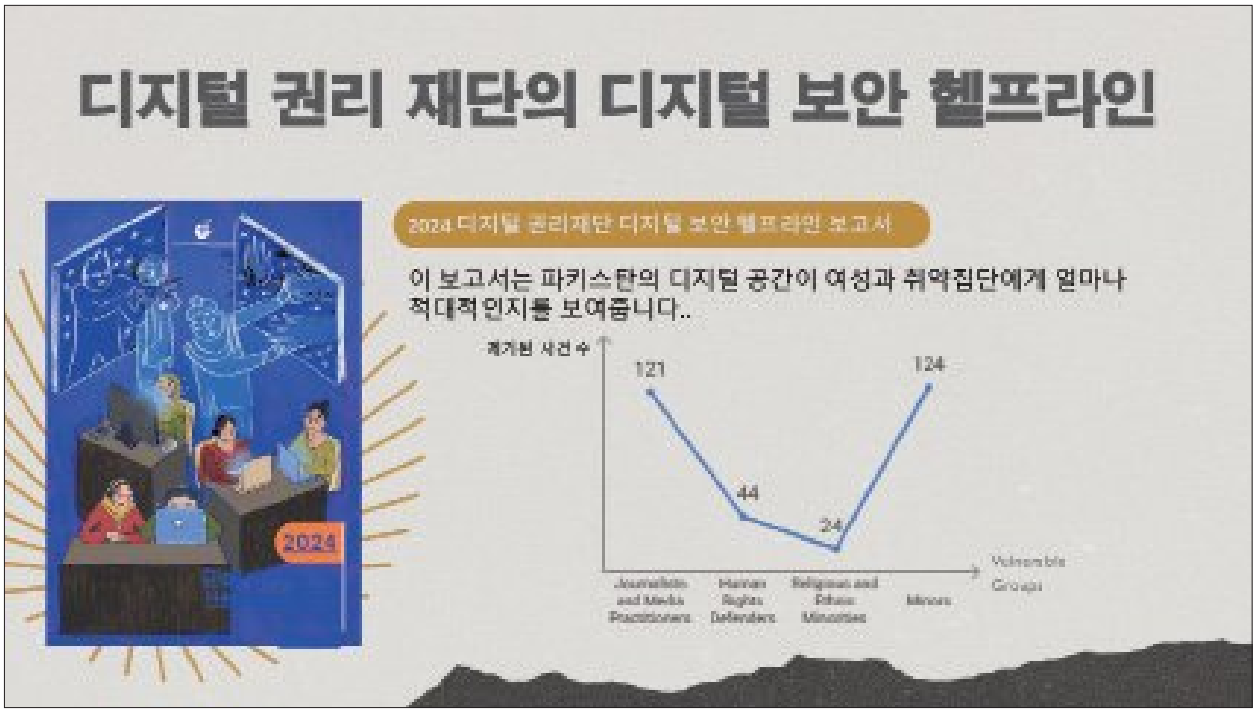

Digital Battlegrounds: Gendered Disinformation, TFGBV, and Hate Speech in the Indo-Pak Escalations
2025. 5. 인도-파키스탄 분쟁 시 디지털 갈등
DRF 조사에 따르면 온라인 게시물 295건 중 73건이 성별 기반 허위정보, 증오 발언, 기술 기반 성폭력을 담고 있었습니다. 강간 협박, 정욕 서사, 드랜스젠더 혐오발언 등이 디지털 전쟁의 무기로 활용되었습니다.
디지털 전쟁터
분쟁은 온라인 공간에서 성별 기반 학대를 심화시키고, 플랫폼의 무대응으로 인해 디지털 공간은 또 다른 전쟁터가 되어 여성과 소수자에게 실제적인 피해를 초래하고 있음을 보여줍니다.


사례


AI 단기 계약 노동(gig) 경제와 노동권
데이터 라벨링 산업
약 170억 달러 규모의 데이터 라벨링 산업은 파키스탄을 포함한 글로벌 사우스의 프리랜서 노동자들에게 과도하게 의존하고 있습니다. 그러나 이들의 노동은 민감하거나 폭력적인 데이터를 다루며, 고된 주석 작업을 요구하는 등 부적절한 조건에 놓여 있습니다.


AI 단기 계약 노동(gig) 경제와 노동권
AI 산업이 이미 막대한 가치를 창출하고 있음에도 불구하고
이와 동시에 등장한 단기 계약 기반 경제는 철저히 가려져 있으며, 노동자들은 기본적인 보호조치 받지 못합니다. 대부분의 계약은 모든AI 같은 대형 기술기업의 하청을 통해 이루어지고, 프리랜서들의 수익은 시간당 1-3달러에 불과합니다. 게다가 마이크로태스크는 파키스탄 고용법상 노동으로 인정되지 않아, 정규 계약이나 법적 권리 보장을 전혀 기대할 수 없습니다.


성별에 따른 디지털 격차
디지털 권리 재단의 연구는 온라인 폭력, 여성의 디지털 공간 배제, 알고리즘 편향이 어떤 집단들을 더욱 악화시킨다는 점을 드러냅니다..
• 오프라인 젠더 기반 폭력은 여성 언론인의 직업적 취약성과 맞물려 나타납니다.
• 여성 언론인을 겨냥한 폭력은 디지털·공적 공간에서 여성의 참여를 줄여 성별 격차를 넘힙니다.
• TFGBV(기술 기반 성별 폭력)는 여성들이 자기검열을 하거나 아예 디지털 플랫폼을 떠나게 만드는 위축 효과를 가져옵니다.
여성이 존재하지만 안전하지 않은 적대적 디지털 환경은 성별화된 디지털 격차를 확대합니다.


디지털 권리 재단의 디지털 보안 헬프라인
2024
2024 디지털 권리재단 디지털 보안 헬프라인 보고서
이 보고서는 파키스탄의 디지털 공간이 여성과 취약집단에게 얼마나 적대적인지를 보여줍니다..
제기된 사건 수
121
44
24
124
Journalists and Media Practitioners
Human Rights Defenders
Religious and Ethnic Minorities
Minors
Vulnerable Groups


디지털 권리 재단의 디지털 보안 헬프라인
2024 디지털 권리재단 디지털 보안 헬프라인 보고서
이 데이터는 여성이 여전히 온라인 괴롭힘의 주된 대상이며, 신고의 대부분을 차지한다는 사실을 보여줍니다.
Gender Breakdown
Male
Female
Non-Binary
Trans Women
Trans Men
No Info
2024


소외 집단의 접근 장벽
정부의 잘못된 우선순위
디지털 공공 인프라와 인권 탈취
통제되지 않은 디지털화와 부실한 데이터 시스템은 소수 집단을 배제하고 차별을 강화함.
고정 브로드밴드 보급률은 1.3%에 불과하며, 5G 도입은 지연되고 있음
성별 및 연금별 격차
정기적으로 인터넷에 접속하는 비율은 남성 53%, 여성 33%에 불과


소외 집단의 접근 장벽
기기 소유와 비용부담
남성의 86%가 핸드폰을 소유하는 반면, 여성의 53%만이 핸드폰을 소유하고 있음.
여성의 안전과 대체 문제
스마트폰 보유 성별 격차는 49%에 이르며, 인터넷 사용 격차는 38%에 이름. 대부분의 여성은 공용폰에 의존하고 있음.
유니세프의 2023 보고서 성별 디지털 격차 줄이기
발루치스탄 지역의 여성과 소녀의 디지털 기술 보유율은 3%미만임.
유니세프의 2023 보고서 성별 디지털 격차 줄이기
신드지방 여성은 기초기술과 디지털 기술에 대해서 뚜렷한 차지에 있음.


정부 주도의 디지털 격차 심화:
분쟁지역에 대한 인터넷 차단, VPN 금지, '국가안보'에 따른 제한
정부 주도의 디지털 격차 심화
분쟁 지역에서 빈번히 발생하는 인터넷 차단 문제. VPN 금지, '국가안보'라는 명분으로 인터넷에 전략적 제한이 강화됨.
신기 관련 법제
2023-24년과 2024 년 동안 기간 동안 발루치스탄에서의 인터넷 차단은 지역 사회와 시민 전여에 부정적인 영향을 미침
새로운 통제 기술
파키스탄은 중국산 인화역 시스템을 사업했으며, 이를 통해 지오정보검출 지역 접속 차단, 이더넷 속도 저하, 표적화된 인터넷 차단을 가능하게 함
디지털 경제에 미치는 영향
지오블로킹과 속도 저하는 안정적이고 빠른 인터넷에 생계를 의존하는 프리랜서와 자영기 계약 노동자들의 활동을 방해함


해결책과
향후 과제
지역 사회의 경험을 글로벌 디지털 정책 포럼과 연결. 인터넷 차단, 성별 기반 피해, 접근 격차 문제를 다루기
지역 및 남남협력
집단 옹호활동을 위해 국가 간 연대 강화
DACSA 남아시아 디지털 책무성 연합
디지털 리터러시 및 안전 교육
지역 사회에 디지털 기술과 온라인 안전에 대한 인식을 높인다.
"Hamara Internet(우리의 인터넷)", "Mahfooz Internet(안전한 인터넷)"
디지털 경제 속 노동자 권리
긱(단기 계약) 노동자와 AI 데이터 노동자들을 위한 공정한 노동 보호 보장
글로벌 기술 기업들의 투명성과 책임성 촉구


감사합니다
DigitalRightsFoundation

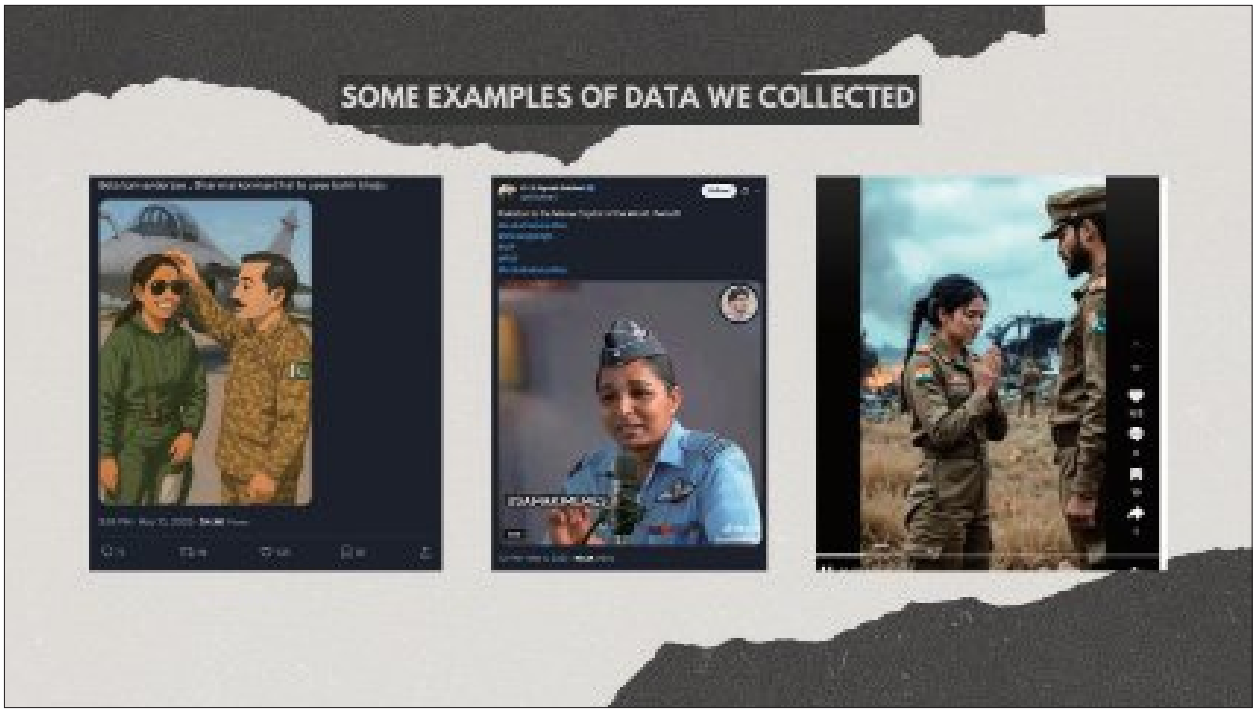
Technological Development and the Digital Divide

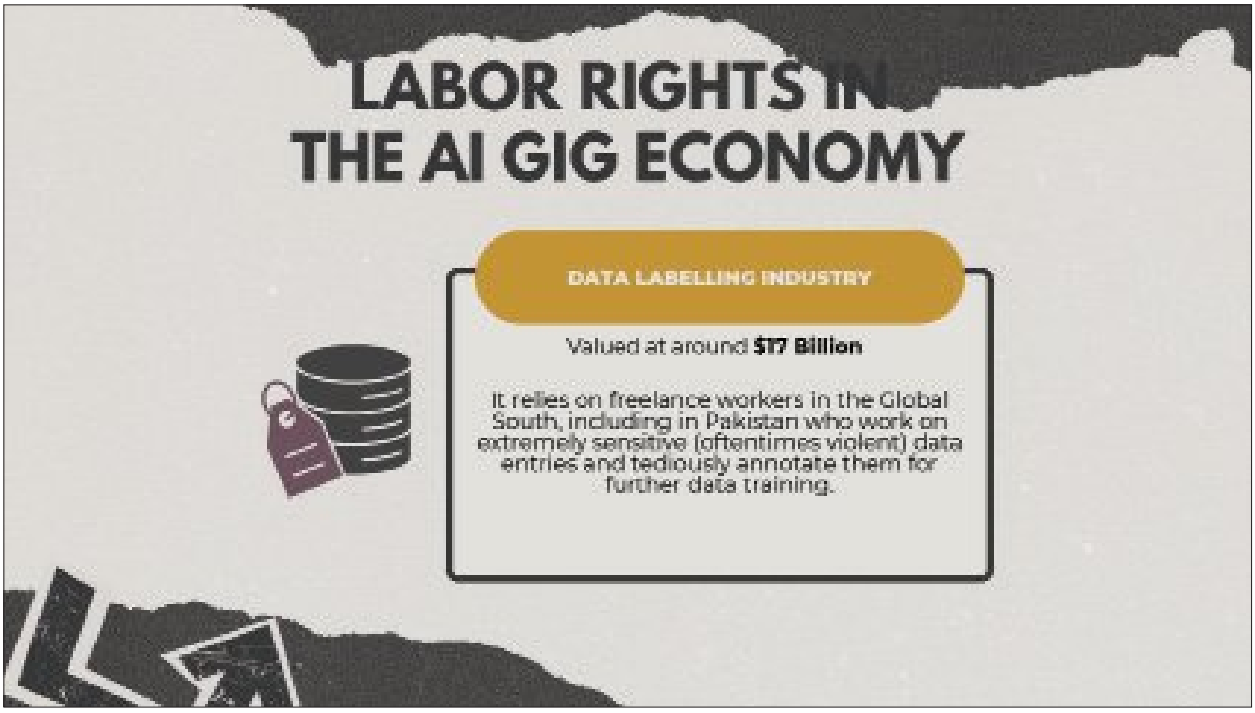
Nighat DAD | Executive Director, Digital Rights Foundation, Member, UN High-level Advisory Body on AI

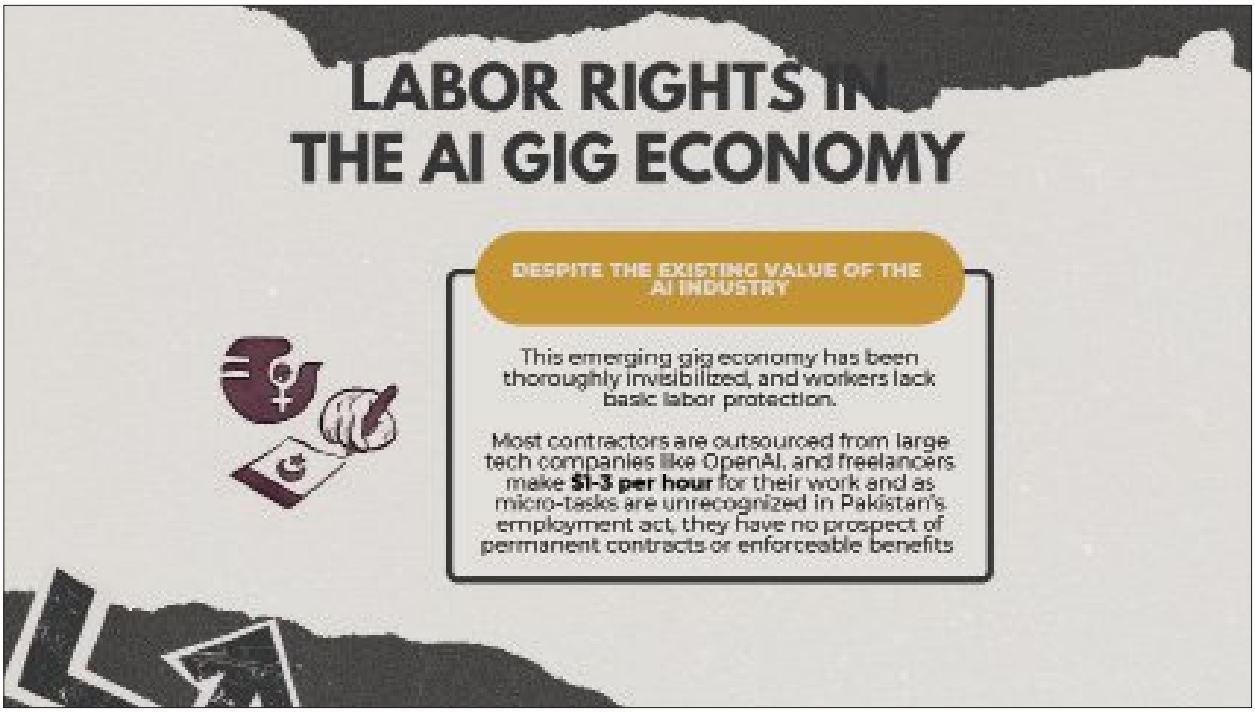

AI-DRIVEN DISINFORMATION AND GENDERED HARMS
THINK ABOUT IT
Generative AI tools are being used to create hyper-realistic deepfakes and manipulate narratives, disproportionately targeting women journalists and public figures.
Gendered AI disinformation campaigns reinforce misogyny and silence women's voices by undermining their credibility, often spilling over into offline harassment and violence.


HONOUR KILLING AND THE AI-FUELLED GLORIFICATION OF GENDER-BASED VIOLENCE ONLINE
NEW TURN IN BALOCHISTAN CASE
The lack of safeguards and transparency in AI systems means women in Pakistan, already facing access and literacy barriers, are especially vulnerable to the new wave of tech-facilitated abuse.


Digital Battlegrounds: Gendered Disinformation, TFGBV, and Hate Speech in the Indo-Pak Escalations
DURING THE MAY 2025 INDIA-PAKISTAN CONFLICT
DRF found that 73 out of a total of 295 online posts about the escalations across platforms carried gendered disinformation, hate speech, or technology-facilitated gender-based violence using rape threats, conquest narratives, and transphobic slurs as weapons of digital warfare.
DIGITAL BATTLEGROUND
This shows how conflict amplifies gendered abuse online, and how platform inaction allows digital spaces to become another front in the battle, with real-world consequences for women and marginalised groups.

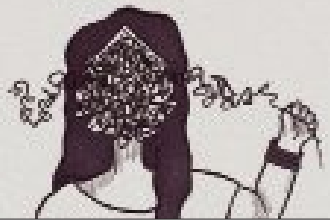

SOME EXAMPLES OF DATA WE COLLECTED


LABOR RIGHTS IN
THE AI GIG ECONOMY
DATA LABELLING INDUSTRY
Valued at around $17 Billion
It relies on freelance workers in the Global South, including in Pakistan who work on extremely sensitive (oftentimes violent) data entries and tediously annotate them for further data training.


LABOR RIGHTS IN
THE AI GIG ECONOMY
DESPITE THE EXISTING VALUE OF THE AI INDUSTRY
This emerging gig economy has been thoroughly invisibilized, and workers lack basic labor protection.
Most contractors are outsourced from large tech companies like OpenAI, and freelancers make $1-3 per hour for their work and as micro-tasks are unrecognized in Pakistan's employment act, they have no prospect of permanent contracts or enforceable benefits

GENDERED DIMENSIONS OF THE DIGITAL DIVIDE

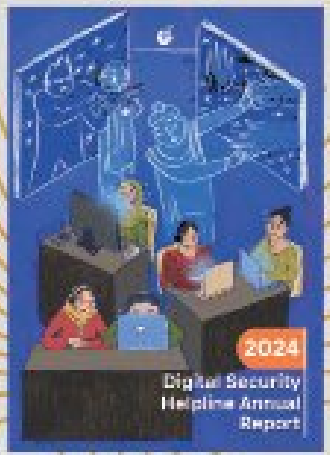
DRF's research on online violence, women's exclusion from digital spaces, and how algorithmic bias compounds these inequalities.



- Offline gender-based violence intersects with professional vulnerabilities of women journalists.
- Violence against women in media deepens the gender gap in participation in digital and public spaces.
- TFGBV creates a chilling effect: women self-censor or leave digital platforms altogether.

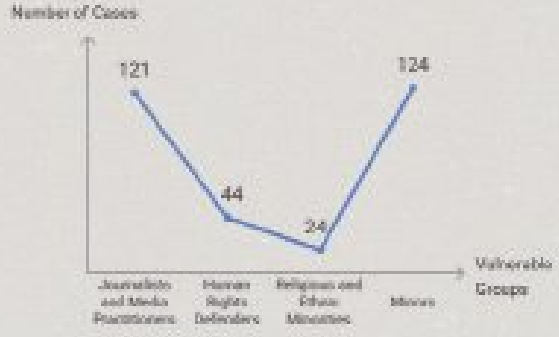
Demonstrates how hostile digital environments contribute to the gendered digital divide, where women are present but unsafe.

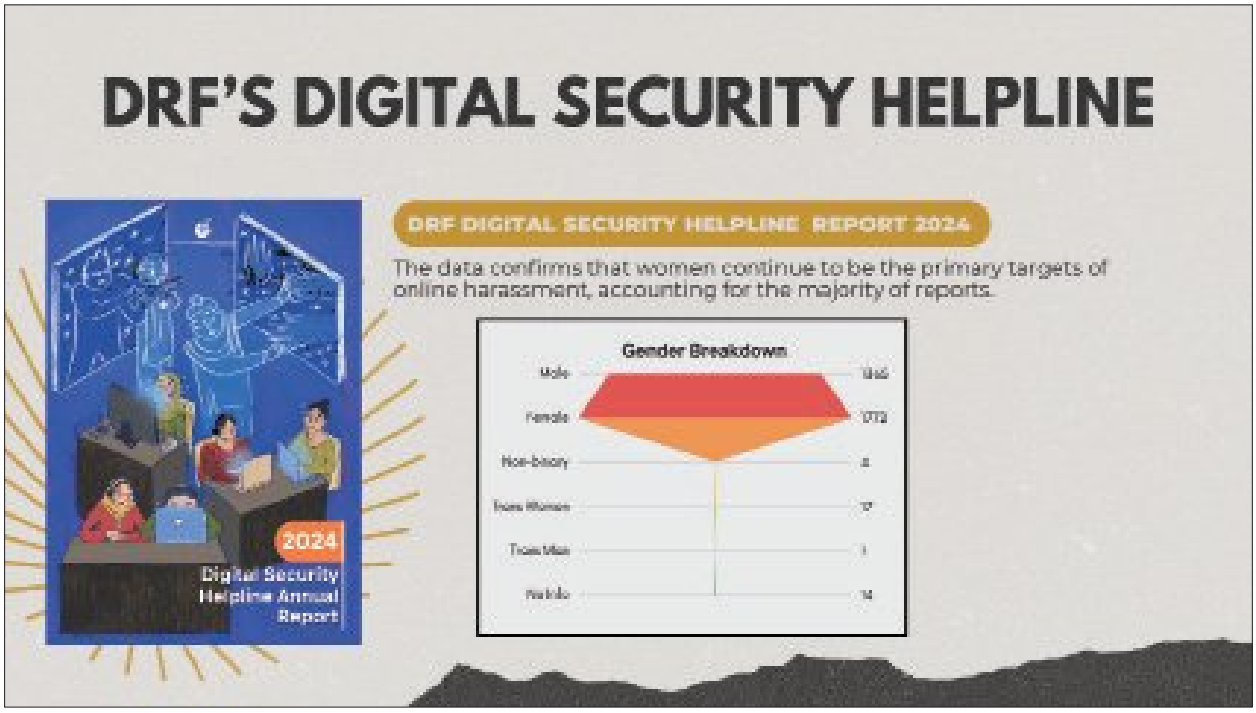
DRF'S DIGITAL SECURITY HELPLINE



DRF DIGITAL SECURITY HELPLINE REPORT 2024

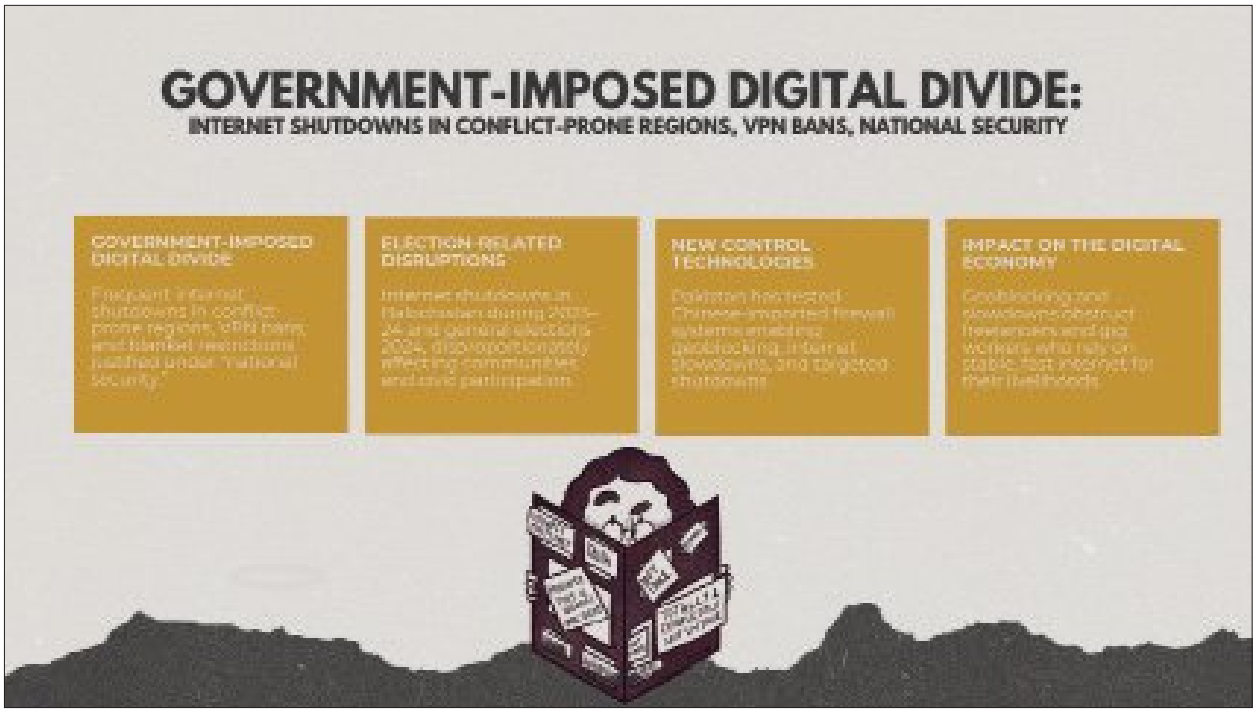
show how hostile Pakistani digital spaces are for women and vulnerable groups

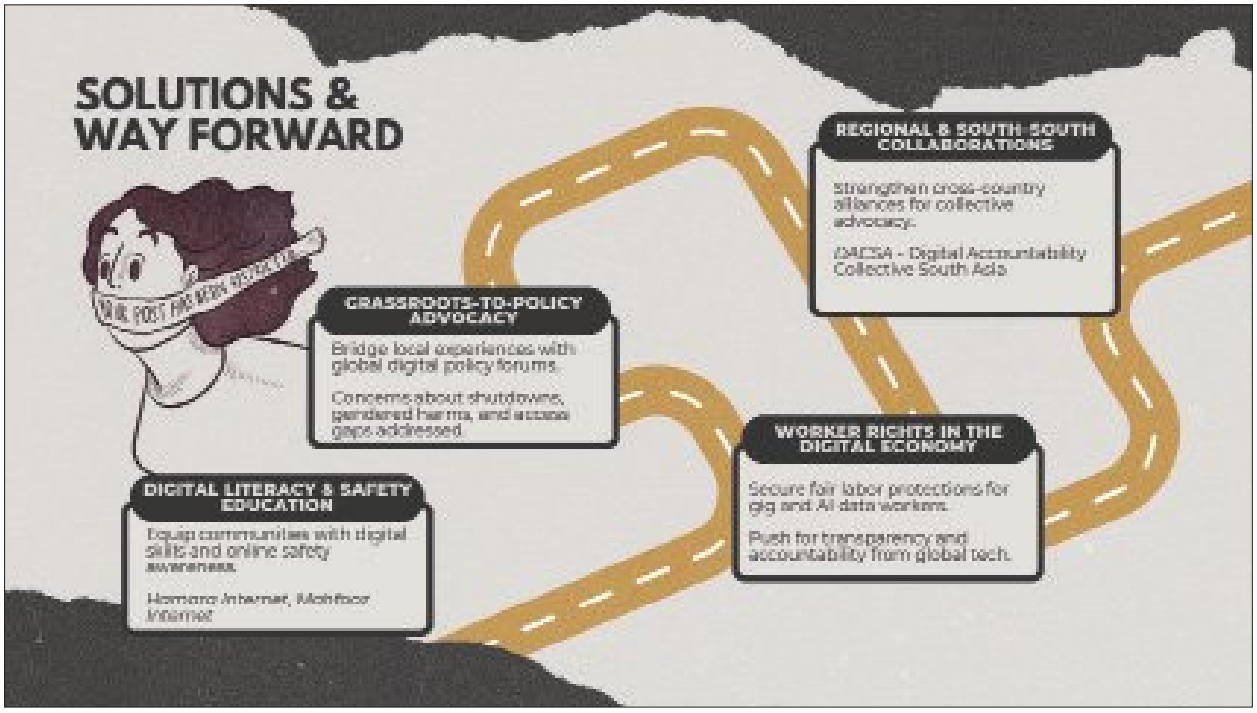



DRF'S DIGITAL SECURITY HELPLINE
2024
Digital Security Helpline Annual Report
DRF DIGITAL SECURITY HELPLINE REPORT 2024
The data confirms that women continue to be the primary targets of online harassment, accounting for the majority of reports.
Gender Breakdown
Male
Female
Non-binary
Trans Women
Trans Men
Not to do


BARRIERS TO ACCESS FOR MARGINALISED GROUPS
2024, PAKISTAN NEW "NATIONAL FIREWALL", COSTING AROUND RS30 BILLION
Government has misdirected its priorities
DIGITAL PUBLIC INFRASTRUCTURE AND HUMAN RIGHTS OP-ED
Unchecked digitalisation and poor data systems, deepen exclusion by pushing marginalised groups
RURAL-URBAN AND GENDERED ACCESS GAPS
Fixed broadband penetration at just 1.3% and major delays in 5G rollout
GENDER AND INCOME-BASED DIVIDES
Only 53% of men and 33% of women regularly access the internet


BARRIERS TO ACCESS FOR MARGINALISED GROUPS
AFFORDABILITY AND DEVICE OWNERSHIP
Mobile phone ownership stands at 86% for men but only 59% for women
SAFETY AND EXCLUSION FOR WOMEN
Gender gap in smartphone ownership is 40% and in internet adoption 38%, with many relying on shared phones.
UNICEF'S 2021 REPORT BRIDGING THE GENDER DIGITAL DIVIDE
Digital skills among young girls & young women aged 15-24 in Balochistan are below 3 %
UNICEF'S 2023 REPORT BRIDGING THE GENDER DIGITAL DIVIDE
In Sindh, females are at a disadvantage in foundational skills and digital skills


GOVERNMENT-IMPOSED DIGITAL DIVIDE:
INTERNET SHUTDOWNS IN CONFLICT-PRONE REGIONS, VPN BANS, NATIONAL SECURITY
GOVERNMENT-IMPOSED DIGITAL DIVIDE
Frequent internet shutdowns in conflict-prone regions, VPN bans, and blanket restrictions justified under "national security."
ELECTION-RELATED DISRUPTIONS
Internet shutdowns in Balochistan during 2023–24 and general elections 2024, disproportionately affecting communities and civic participation.
NEW CONTROL TECHNOLOGIES
Pakistan has tested Chinese-imported firewall systems enabling geoblocking, internet slowdowns, and targeted shutdowns.
IMPACT ON THE DIGITAL ECONOMY
Geoblocking and slowdowns obstruct freelancers and gig workers who rely on stable, fast internet for their livelihoods.


SOLUTIONS &
WAY FORWARD
REGIONAL & SOUTH-SOUTH COLLABORATIONS
Strengthen cross-country alliances for collective advocacy.
DACSA - Digital Accountability Collective South Asia
GRASSROOTS-TO-POLICY ADVOCACY
Bridge local experiences with global digital policy forums.
Concerns about shutdowns, gendered harms, and access gaps addressed
WORKER RIGHTS IN THE DIGITAL ECONOMY
Secure fair labor protections for gig and AI data workers.
Push for transparency and accountability from global tech.
DIGITAL LITERACY & SAFETY EDUCATION
Equip communities with digital skills and online safety awareness.


THANK
YOU
DRF
DigitalRightsFoundation

[발표 4 | Speaker 4]

생존을 위한 시대, AI의 역할

AI in the Age of Survival

마이클 키웻

Michael KWET

요하네스버그대학교 사회변화연구센터 선임연구원, 예일대학교 방문 교수

Senior Researcher, University of Johannesburg

Visiting Professor, Yale University

생존을 위한 시대, AI의 역할

마이클 키웻 | 요하네스버그대학교 사회변화연구센터 선임연구원, 예일대학교 방문 교수

여러분, 안녕하십니까? 오늘 이 자리에 함께하게 되어 진심으로 영광입니다. 이번 회의의 주제는 매우 시의적절하다고 생각합니다. 저는 오늘, 인공지능과 인권의 연관성을 이해하기 위해 반드시 전면에 두어야 할 두 가지 핵심 의제가 있다고 말씀드리고자 합니다. 그것은 바로 환경적 지속가능성, 즉 전 세계 자원 이용의 한계 문제와, 그리고 디지털 식민주의, 다시 말해 미국이 전 세계 디지털 생태계를 일극적으로 지배하고 있는 현실입니다.

인공지능과 지구 디지털 경제에 대해서 이야기하기 전에, 먼저 환경적 지속가능성에 대해서 말씀드리겠습니다.

유엔을 비롯한 여러 기관에 따르면, 우리는 현재 기후변화, 생물다양성 손실, 오염이라는 삼중 위기에 직면해 있습니다. 아마존 열대우림 파괴, 전 세계 해양 생물의 4분의 1을 품고 있는 산호초의 집단 붕괴, 빙하의 급속한 융해, 대서양 해류(AMOC)의 붕괴 등과 같은 돌이킬 수 없는 환경적 전환점에 가까워지고 있습니다. 이러한 사건이 발생한다면 인류를 포함한 지구 생명체에 치명적인 결과를 초래할 것입니다. 대부분의 과학자들은 산업화 이전 대비 지구 평균기온이 1.5도 상승할 경우 주요 전환점이 촉발될 가능성이 있으며, 현재 추세라면 2도 상승에 도달하게 될 것이고 그 가능성은 "매우 높다"고 말합니다.

재생에너지로의 전환이 진행 중이지만, 경제성장이 계속될수록 지구 과열을 막기는 더 어려워집니다. 새로운 기술이 발명되고 대규모로 확산되지 않는 한, 경제 성장 자체가 재생에너지 효과를 상쇄해 지구 온난화를 가속화할 것이라는 전망이 지배적입니다.

또한 인간의 경제활동은 기후변화 뿐만 아니라 전례 없는 속도로 생물다양성을 파괴하고 있습니다. 많은 과학자들은 인류가 "제6차 대멸종"을 일으키고 있다고 경고하며, 야생 동물 개체의 수를 급감시키고 있다고 주장합니다. 세계자연기금(WWF)에 따르면, 1970~2020년 사이, 5,495종 척추동물의 개체 수가 평균 73% 감소했으며, 곤충은 매년 약 2%씩 줄어들고 있습니다. 이 속도라면 40년안에 모든 곤충의 절반이 사라질 것입니다. 미국에서는 1970년대 이후 지난 50여 년간 조류 개체의 약 3분의 1이 사라졌습니다. 우리는 지구 곳곳에서 생명을 위협하며 서식지를 파괴하고 있는 것입니다. 현재 생물다양성 손실의 주된 원인은 기후변화가 아니라 인간의 경제 활동입니다. 이는 곧, 설령 기후과열이 문제가 아니더라도, 무한한 경제 성장은 그 자체로 지속 불가능하다는 뜻입니다.

이러한 배경에서, 과학자와 정책 전문가의 71%는 세계적 차원의 경제 성장이 지속 불가능하다고 답하고 있습니다.

이것은 매우 중요한 지점입니다. 일반적으로 자본주의 체제 옹호자들은 "성장은 빈부격차를 만들지만, 충분한 성장이 이루어지면 결국 전 세계 빈곤을 퇴치할 수 있다"고 주장합니다. 그러나 물질흐름학(material flows) 연구자들의 자원 사용에 대한 체계적 분석은 전혀 다른 결론을 보여줍니다. 지구적 위험을 최소화하기 위해서는 전 세계 원자재 사용량을 절반으로 줄이고 상한선을 설정해야 한다는 것입니다. 이는 자본주의와 양립할 수 없습니다. 자본주의는 더 많은 생산을 위한 투자로 경제를 운영하며, 기하급수적으로 성장하는 구조를 갖고 있습니다. 기하급수적 성장을 이해하는 사람이라면, 시간이 지남에 따라 그 속도가 급격히 빨라진다는 것을 잘 알 것입니다.

또한 이는 빈곤 퇴치와 사회 정의와도 양립하기 어렵습니다. 현재 세계 경제는 연간 120조 달러를 생산하며, 이를 1인당 생산량으로 환산하면 연간 약 2만4천 달러에 해당합니다. 그러나 생산물은 국가 간·국가 내에서 심각한 수준으로 불평등하게 분배됩니다. 예컨대 미국은 전 세계 인구의 4%에 불과하지만, 세계 부의 31%, 금융자산의 45%를 소유하고 있습니다.

데이터를 면밀히 살펴보면, 모든 인류가 지구적 한계 안에서 서구식 중산층 생활을 누리려면 부와 소득을 국가 간·국가 내에서 공평하게 재분배해야 한다는 결론에 도달할 수밖에 없습니다. 유엔은 이러한 과정을 "수축과 수렴(contraction and convergence)"이라고 부릅니다. 이는 전 세계 자원 사용량을 줄이고, 부유한 나라와 가난한 나라 사이에 자원 소유가 점차 균등해지는 것을 의미합니다. 저는 이를 "정의로운 탈성장 전환(just degrowth transition)"이라 부릅니다. 이는 재생에너지 전환과 일부 산업형 농업 규제만을 요구하는 자본주의식 "정의로운 전환(just transition)"과는 본질적으로 다릅니다.

역사적 발전 과정을 고려하면, '수축과 수렴(contraction and convergence)'의 필요성은 더욱 설득력을 갖습니다. 환경경제학에서는 두 가지 주요한 부채 개념을 다룹니다. 첫째는 기후부채(climate debt)로, 산업혁명 이후 현재까지 각국이 배출한 탄소량을 합산해 산출합니다. 이 기준에 따르면, 미국은 전 세계 초과배출의 40%를 차지하며, EU 28개국은 29%로 그 뒤를 잇습니다. 반면 글로벌 사우스(Global South)은 단 8%에 불과합니다.

둘째는 생태부채(ecological debt)로, 국가별 과잉 자원 사용량을 계산한 것입니다. 미국은 27%, EU 28개국은 25%, 중국은 15%를 차지하며, 나머지 글로벌 사우스(Global South)는 8%에 머물고 있습니다.

미국은 전 세계 인구의 4%에 불과하지만, 세계 부의 31%와 금융자산의 45%를 보유하고 있습니다. 이는 미국이 현재 전 세계 자원을 과도하게 소비하고 있을 뿐 아니라, 수십 년 동안 도로, 사무실, 주택과 같은 핵심 인프라를 건설하며 자원을 과잉 사용해 번영을 쌓아온 데 대해서도 세계에 환경적 부채를 지고 있음을 보여줍니다. '탈성장(degrowth)'의 관점에서 볼 때, 미국에 집중된 이 부와 자원 소비는 결코 지

속 가능하지 않습니다. 이제는 미국이 환경적 부채를 상환하고, 세계가 공유해야 할 한정된 자원을 자신의 몫 이상 소비하지 않도록 요구하는 세계적 차원의 운동이 필요합니다.

이러한 맥락 속에서 우리는 디지털 기술과 인권을 이해해야 합니다. 제가 집필한 『디지털 탈성장: 생존의 시대에 기술을 바라보다(Digital Degrowth: Technology in the Age of Survival)』는 전 세계 디지털 경제의 소유 구조를 체계적으로 집계한 최초의 연구였습니다. 이를 위해 저는 초국적 기술기업의 국가별 소유 현황, 스타트업 부문, 투자, 연구개발, 지적재산권뿐 아니라 반도체, 클라우드 컴퓨팅, 인공지능, 빅데이터, 운영체제, 하드드라이브, 이메일, 스트리밍 서비스 등 다양한 제품과 서비스까지 분석했습니다. 흔히 세계 디지털 경제가 미국과 중국이라는 양극 체제로 이루어져 있다고 생각하지만, 이는 전혀 사실과 다릅니다. 실제로는 미국이 전 세계 디지털 경제를 일극적으로 지배하고 있습니다.

미국의 디지털 분야 패권이 얼마나 압도적인지를 보여주기 위해 몇 가지 사실만 말씀드리겠습니다.

- 2023년 11월 기준, 상위 943개 기술 기업 가운데
- 미국은 전체 기업의 55%, 시가총액의 77%, 매출의 59%를 차지합니다.
- 이에 비해 중국은 기업 수 6%, 시가총액 6%, 매출 11%에 불과합니다.
- 그러나 이마저도 실제 상황을 다 보여주지 못합니다. 중국 기업의 대부분은 자국 내에서만 영업 활동을 하고 있기 때문입니다.
- 상위 15개 중국 기술기업의 경우, 해외 매출 비중은 19%에 그칩니다.
- 반면 상위 15개 미국 기술기업은 매출의 49%를 해외에서 올리고 있습니다.

제 책에 나와 있듯이, 다른 디지털 경제 상황도 이와 동일합니다.

물질적 자원의 한계를 고려하면, 부의 집중은 더 이상 환경적 지속가능성과 양립할 수 없습니다. 모든 이들이 나눠 쓸 만큼 자원이 충분하지 않기 때문입니다. 여기서 첫 번째 중대한 문제가 드러납니다. 미국이 여러 지표에서 가장 수익성이 높은 전 세계 경제 부문, 즉 디지털 부문을 소유하고 지배하고 있다는 점입니다. 세계 디지털 경제는 학자들이 말하는 '생태적으로 불평등한 교환' 구조를 강화합니다. 즉, 미국을 중심으로 한 부유한 국가들은 계산 능력과 지식을 소유·통제하며 고차원적 "사고"를 담당하는 반면, 가난한 국가들은 땅을 파서 광물을 캐고, 상품작물을 생산하며, 소셜미디어에서 유해 콘텐츠를 걸러내거나 AI를 위한 데이터 라벨링 작업 등 단순 노동을 떠맡습니다. 여기에 더해 글로벌 노스는(중국을 포함하여) 자국의 천연자원을 고갈시키고 있으며, 글로벌 사우스는 글로벌 공급망의 최하위 단계에서 발생하는 환경적 피해를 감당하고 있습니다.

결국 세계 디지털 경제는 생태적으로 불평등한 교환 구조를 강화합니다. 미국을 비롯한 글로벌 노스는

첨단 기술과 지식의 소유권을 독점하고, 글로벌 사우스는 광물 채굴, 농산물 생산, 소셜미디어 콘텐츠 정화와 데이터 라벨링과 같은 저임금 노동을 떠맡고 있습니다. 이를 통해 미국과 북반구 엘리트는 막대한 부를 축적하며, 미국의 기술 노동자들조차 평균 연봉 30만 달러를 받으며 혜택을 누리고 있습니다.

이러한 생태적으로 불평등한 교환은 미국의 부유층에게 불균형적으로 이익을 가져다줍니다. 미국의 기술 기업 경영자와 투자자들은 수십억 달러, 수백만 달러의 자산을 가진 거부가 되었고, 화이트칼라 기술 노동자들 또한 그 혜택을 누리고 있습니다. 이들의 연평균 임금은 약 30만 달러로, 이는 지구 자원의 공정하고 지속가능한 몫을 훨씬 넘어서는 수준입니다.

'디지털 탈성장(digital degrowth)'의 관점에서 우리는 사실을 분명히 할 필요가 있습니다. 전 세계 디지털 경제의 중심에는 미국 제국이 존재한다는 것입니다. 그러나 주류 학자들과 언론인들, 대부분 미국 출신인 이들에 의해 이러한 사실이 이야기에서 지워진 것은 결코 우연이 아닙니다. 이들 지식인들이 하버드, 예일 같은 명문 대학이나 마이크로소프트, 구글, 그리고 거대 재단들과 같은 수십억 달러 규모 기관으로부터 매년 수십만 달러의 지원을 받는 것 역시 우연이 아닙니다. "미국 제국"이라는 개념은 그들의 저작에서 완전히 빠져 있거나, 있더라도 단 한 단락으로 축소되어 마치 대수롭지 않은 것처럼 다뤄집니다. 마치 19세기 영국 학자들이 "사회 정의"를 논하면서도 영국 제국의 존재를 언급하지 않거나, 책 속의 몇 줄로만 처리한 것과 다름없습니다. 오늘날 디지털 정치와 인권을 둘러싼 세계적 담론이 바로 그러한 상황에 놓여 있습니다.

지난 몇 년 동안 학자들과 정책입안자들은 디지털 기술과 환경의 교차 지점에 주목해 왔습니다. 특히 에너지 소비, 탄소 배출, 물 사용량, 그리고 그보다는 덜하지만 지역 생태계 오염에 집중했습니다. 물론 이것은 중요한 문제이지만, 이러한 통념은 매우 왜곡되어 있습니다. 대중 언론, 비정부기구, 그리고 주요 학자들의 논의를 접하다 보면, 마치 인공지능이 이제 기후변화와 수자원 고갈의 주된 원인인 것처럼 보입니다. 그러나 실제 사실은 전혀 다른 이야기를 들려줍니다.

제 책의 한 챕터인 "탈식민지화로서의 디지털 탈성장(Digital Degrowth as Decolonization)"의 계산에 따르면, 다음과 같은 점을 확인할 수 있습니다.

AI는 데이터센터 용량의 약 20%를 사용하는 것으로 알려져 있습니다. 따라서 데이터센터를 통해 AI가 정보통신기술(ICT) 부문 전체 온실가스 배출에서 차지하는 비중은 약 2.5%에 불과하며(전 세계 배출량의 0.035%), ICT 부문 전력 사용량에서 차지하는 비중도 약 4.9%(전 세계 전력 사용의 0.2%)에 그칩니다. 설령 학계 연구가 ICT 부문의 전력 사용과 배출량을 다소 과소평가했더라도, 그리고 현재 진행 중인 AI·데이터센터 '붐'으로 인해 제시된 최고치 전망을 받아들인다 하더라도, 데이터센터와 AI의 전 세계적 배출 및 전력 사용량은 화석연료(현재 전 세계 온실가스 배출의 68%를 차지)나 식량 체계(전 세계 온실가스 배출의 약 3분의 1을 차지)에 비하면 극히 미미한 수준입니다.데이터센터의 에너지 및 물 사용은 전 지구적 차원에서는 매우 작은 환경 발자국을 남기며, 영향은 주로 지역적이거나 일부 경우에는 국가적 수준에서 더 뚜렷하게 나타납니다.

AI와 환경에 관한 대중적 논의는 본질을 보지 못하고 지엽적인 문제에 매달리고 있습니다. 디지털 부문이 생태 파괴에 영향을 미치는 핵심 요인은, 전 세계적 차원에서 보면 지극히 미미한 수준에 불과한 AI와 데이터센터의 직접적 환경 발자국, 혹은 ICT 부문 전체의 배출이 아닙니다. 진짜 문제는 엘리트 축적, 생태적으로 불평등한 교환, 자본주의의 디지털화, 그리고 미국 제국의 강화가 정의로운 탈성장 전환을 가로막고 있다는 사실입니다. 이것이야말로 기술정치를 이해하는 데 있어 핵심적인 지점이지만, 제가 책에서 보여주었듯 디지털 경제와 환경의 관계를 다룬 기존 문헌들은 모두 이를 놓치고 있습니다.

미국은 디지털 경제의 핵심 축인 하드웨어, 소프트웨어, 네트워크 연결을 소유하고 통제하기 때문에, 정보의 흐름을 추출하고 재편하며, 온라인 공간을 지배하고, 다른 나라들을 협상 끝에 굴복시킬 수 있습니다. 경제 영역을 넘어, 미국은 인공지능과 같은 첨단 디지털 기술을 해외에 위치한 800여 개 군사 기지에 통합하고 있는데, 이는 전 세계적인 지구 온난화와 생물다양성 손실을 직접적으로 야기할 뿐 아니라, 미국이 세계 경제에 대한 지배력을 유지하고 현 상태를 지속하는 데 사용됩니다. 만약 어떤 나라가 미국의 패권에 진지하게 도전한다면, 무역 제재와 군사적 보복에 직면할 것입니다.

이는 다음 질문으로 연결됩니다. 그렇다면 우리는 무엇을 할 수 있을까요? 무엇이 문제인지 제대로 이해하지 못하면 해결할 수 없습니다. 따라서 우리가 취해야 할 첫걸음은 디지털 정치의 틀을 '디지털 탈성장'의 관점에서 새롭게 세우고, 디지털 식민주의와 자본주의에 맞서는 것입니다. 여기에는 중심적 책임자로서의 미국에 대한 반대 뿐 아니라, 다른 행위자들이 저지르는 해악에 대한 반대도 포함됩니다. 그 예로는 콩고에서 코발트 채굴 과정에서 아프리카 광부들을 착취하는 중국, 그리고 세계 각국에서 엘리트 축적을 노리며 부를 얻으려는 실리콘밸리식 모방 기업들을 들 수 있습니다.

환경적 재앙을 막을 수 있는 시간이 얼마 남지 않은 만큼, 우리는 삶의 방식을 신속히 전환해야 합니다. 그리하여 우리가 오래전에 이루었어야 한다고 모두가 알고 있는 것, 바로 자연과 조화 속에서 서로 간의 완전한 평등을 실현해야 합니다. 체계적인 전환에는 체계적인 계획이 필요합니다. 이러한 맥락에서 저는 생태사회주의적 '디지털 기술 협약'을 제안했습니다. 세부 사항을 길게 논의할 시간은 많지 않지만, 여기에는 '정의로운 탈성장 전환'이 포함됩니다. 즉, 인공지능, 클라우드 컴퓨팅, 온라인 플랫폼을 비롯한 계산과 지식의 수단을 착취적인 고용주나 권위적인 관료가 지휘하는 것이 아니라, 노동자와 지역 공동체의 손에 직접 맡기는 것을 의미합니다. 만약 이것이 평등한 세상을 갈망하는 이들의 비현실적인 꿈처럼 들린다면, 지배와 착취, 끝없는 성장을 이어가면서도 우리의 유일한 지구를 파괴하지 않을 수 있다고 생각하는 것은 그보다 더 비현실적입니다.간단히 말해, 소수 엘리트의 축적과 착취가 아니라 나눔과 조화를 제도화해야 합니다.

요컨대, 우리는 대대적인 변화를 필요로 하는 새로운 시대에 살고 있습니다. 우리가 직면한 위기의 심각성에 대해 솔직히 이야기하고, 진정한 긴박감을 가지고 행동해야 할 때입니다.

감사합니다. 질문을 기대하겠습니다.

AI in the Age of Survival

Michael KWET | Senior Researcher, University of Johannesburg, Visiting Professor, Yale University

Hi. I am truly honored to be here today, and the subject matter of this conference couldn't be more timely. Today I will argue that if we want to understand the connection between artificial intelligence and human rights, we have to bring two key issues to the fore: environmental sustainability, which includes limitations on global material resource use, and digital colonialism, whereby the United States exercises unipolar dominance over the global digital ecosystem.

Before we get to AI and the global digital economy, let's start with environmental sustainability.

According to the United Nations, and others, we face a triple crisis of climate change, biodiversity loss, and pollution. We should add to this that we are dangerously close to triggering several environmental tipping points – irreversible events like the destruction of the Amazon rain forest, the mass die-off of coral reefs (which house ¼ of all ocean life), the melting of glacial ice sheets, and the collapse of the AMOC ocean current – which, if they occur, will do catastrophic damage to life on the planet, including us. According to most scientists, we are approaching 1.5C above the preindustrial level – at which point, they warn, the likelihood of triggering major tipping points is "possible". On the current trajectory, we will hit 2C above the preindustrial level, and the probability of triggering these tipping points becomes "likely".

As we are trying to transition to renewable energy, the more we grow our economies, the harder it becomes to stop overheating the planet. Barring the invention and mass-scaling up of new technologies, projections show that worldwide economic growth will cancel out the effects of renewables and keep increasing the global temperature.

In addition to overheating the planet, human economic activity is driving biodiversity loss at an alarming pace. Many scientists argue that humans are causing a "sixth extinction" event, and we are also causing wildlife populations to plummet. According to the World Wildlife Foundation, the average size of vertebrate species has fallen by an eye-watering 73%

among 5,495 vertebrate species sampled between 1970 and 2020. Entomologists estimate that insect loss is about 2% per year – at that rate, one half of all insects will be gone within four decades. Almost one third of the US bird population has disappeared from the skies since 1970. We're basically killing off life and destroying habitats everywhere we go. For the time being, the primary cause of biodiversity loss is human economic activity, not climate change, meaning that even if overheating the planet weren't an issue, economic growth would still be unsustainable because we're over-exploiting nature.

For this reason, a staggering 71% of scientists and policy experts now believe that worldwide economic growth is unsustainable.

This is a crucial point. Most people who accept the capitalist system argue that yes, capitalism does create rich and poor, but with enough growth, we can eventually eradicate global poverty. But material flows scientists have evaluated material resource use systematically, and they've concluded that for a low-risk environmental scenario, global raw material resource use should be cut by one half and capped. This is incompatible with capitalism, which plans the economy through investment in that which produces more and more and grows at an exponential rate. And those familiar with exponential growth understand that growth accelerates dramatically over time.

It is also incompatible with poverty alleviation and social justice. On the global scale, the world economy produces $120 trillion per year which, divided by eight billion people, produces about $24,000 per head per year. But the fruits of economic production are concentrated both between and within countries. At the global level, the United States has 4% of the population, yet 31% of the wealth and 45% of the financial assets. Within countries, economic inequality is also highly inequitable. If you look closely at the data, then one is forced to conclude that in order to provide a Westernized middle class standard of living to all humans and stay within planetary limits, wealth and income needs to be distributed equally between and within countries. The United Nations calls this process "contraction and convergence", whereby global material resource use contracts and the ownership of resource converges between rich and poor. I call this a "just degrowth transition", which is very different from a capitalist "just transition" that only requires a transition to renewable energy and some regulations on industrial agriculture.

The case for contraction and convergence is even stronger when we consider historical development. In environmental economics, there's two primary forms of debt. The first is called climate debt. This is tabulated by calculating how much carbon each country has

emitted from the industrial revolution to the present. The United States is responsible for 40% of national overshoot, with the EU-28 following closely at 29%. The Global South is responsible for just 8%.

There's another debt, called ecological debt, which tabulates excess material use by country. The United States is responsible for 27%, followed by the EU-28 at 25%. China accounts for 15% and the rest of the Global South 8%.

Here we can see the United States, with 4% of the world's population, 31% of the world's wealth and 45% of the financial assets, not only presently over-consumes the world's material resources, but it owes the rest of the world a debt for its over-consumption of the resources it consumed over decades to build its prosperity, including critical infrastructure like roads, offices, and houses. From a degrowth perspective, wealth concentration into the hands of Americans is unsustainable. There needs to be a global movement demanding that the US pay its environmental debts and stop consuming more that its fair share of the world's finite resources.

It's within this context that we should understand digital technology and human rights. My book, Digital Degrowth: Technology in the Age of Survival, was the first study to systematically tabulate who owns the global digital economy. To do this, I assessed national ownership of transnational tech corporations, the startup sector, investment, research and development, intellectual property, as well as products and services like semiconductors, cloud computing, artificial intelligence, big data, operating systems, hard drives, email, streaming entertainment, and more. The widespread belief that there are two poles in the global digital economy – the United States on one end, and China on the other – is wildly incorrect. In reality, the United States has unipolar dominance over the global digital economy.

Let me state just a few facts here, just to show the extreme extend of US hegemony in the digital sphere:

- As of November 2023, of the top 943 tech corporations:
- The United States owns 55% of the companies, 77% of the market cap, and 59% of the revenue.
- China, by comparison, has 6% of the companies, 6% of the market cap, and 11% of the revenue.
- Even this understates the case: most Chinese corporations do business within mainland China.
- Of the top 15 Chinese tech corporations, only 19% of the revenue comes from abroad.

• The top 15 US tech corporations, by comparison, get 49% of their revenue from overseas.

The rest of the digital economy looks the same – it's all documented in the book.

Now if you factor in material resource limitations, wealth concentration is no longer compatible with environmental sustainability. There simply isn't that much to go around. Here we can see the first major problem: the United States owns and controls what is, by many metrics, the most lucrative part of the global economy: the digital sector. The global digital economy reinforces what scholars call ecologically unequal exchange, whereby the rich, led by the United States, own and control the means of computation and knowledge, and perform the higher-level "thinking", while the poor dig in the dirt for minerals, produce cash crops, and carry out menial tasks like cleansing social media feeds of disturbing content or labeling data for AI. The Global North, joined by China, is also depleting their natural resources while the South suffers from environmental externalities imposed on the bottom of the global supply chain.

This ecologically unequal exchange disproportionately benefits wealthy Americans, whose tech executives and investors have become multi-billionaires and multi-millionaires. White collar tech workers are also beneficiaries, with average salaries $300,000 per year, which comprises much more than their fair and sustainable share of the planet's wealth.

So from a digital degrowth perspective, we need to be clear about the facts: there is a US empire at the center of the global digital economy. Yet it is no coincidence that it has been written out of the story by the leading scholars and journalists, who are mostly from the United States. It is also no coincidence that these same intellectuals receive hundreds of thousands of dollars per year from multi-billion dollar institutions like Harvard, Yale, Microsoft, Google, and the rich foundations. The notion of an American Empire is either completely absent from their publications or reduced a paragraph, as if it is insignificant. Imagine 19th century British scholars writing about "social justice" as if the British Empire doesn't exist, or reducing it to a few lines in a book. That's where we are today in the global conversation about digital politics and human rights.

Over the past few years, intellectuals and policymakers have turned attention to the intersection between digital technology and the environment, with a specific focus on energy consumption, carbon emissions, water consumption, and to a lesser degree, pollution to local habitats. While this is certainly an issue, the story told is wildly misleading. Reading the popular press, non-governmental agencies, and leading scholars, you would think that artificial

intelligence is now a major cause of climate change and water depletion. Yet the facts tell a different story.

As I put it in a book chapter, "Digital Degrowth as Decolonization", I provided calculations:

AI reportedly uses about 20 per cent of data-centre capacity; thus, via data centres, it may contribute as little as 2.5 per cent of GHG emissions within the ICT sector itself (or 0.035 per cent of global emissions) and 4.9 per cent of ICT-sector electricity use (or 0.2 per cent of global electricity use). Even if scholarly findings somewhat understate ICT-sector electricity use and emissions, and even if we accept the high-end projections due to the present AI/data centre 'boom', global emissions and electricity use from data centres and AI are tiny by comparison to fossil fuels (with CO_2 accounting for 68 per cent of current GHG emissions) and the food system (accounting for about one-third of all GHG emissions). Energy and water consumption by data centres creates a very small environmental footprint on a global scale; impacts are more acute at the local or, in a few cases, national level.

The popular conversation about AI and the environment misses the forest for the trees. The central contribution of the digital sector to ecocide is not the "direct" environmental footprints of AI and data centres – which is tiny on a global scale – or even the ICT sector as a whole. Rather, it is the elite accumulation, ecologically unequal exchange, digitalization of capitalism, and strengthening of the American Empire that precludes a just degrowth transition. This point is central to understanding tech politics, yet it has been missed by all the literature on the relationship between the digital economy and the environment, as I've illustrated in the book.

Because the United States owns and controls the core pillars of the digital economy – the hardware, software, and network connectivity – it is able to extract and shape the flow of information, dominate the online space, and bargain countries into submission. Beyond the economic domain, it also integrates advanced digital technology, such as artificial intelligence, into its global military force of 800 bases on foreign soil, which not only directly contributes to global heating and biodiversity loss, but is used to maintain its dominance over the global economy and keep the status quo intact. If countries begin to seriously challenge US supremacy, they face the prospect of trade sanctions and military backlash.

This leads us to the next question: what can we do about it? Well, you can't fix something you don't understand, so the first step is to reframe digital politics around digital degrowth, with explicit opposition to digital colonialism and capitalism. This includes opposition to the United States as the central culprit at the core, as well as harms carried out by others, such as Chinese exploitation of African miners for cobalt in the Congo and Silicon Valley-style knock-offs trying to strike it rich for elite accumulation within countries across the world.

Given the small window left to avert environmental catastrophe, we need a rapid transformation of our way of life to produce what we all know should have been done a long time ago – full equality with each other in harmony with nature. Systemic transformation requires systemic planning, and to this effect I've proposed an ecosocialist Digital Tech Deal. While there isn't much time to discuss the details, this would entail a just degrowth transition that would place the means of computation and knowledge – including AI, cloud computing, and online platforms – directly into the hands of workers and communities, without extractive bosses or authoritarian bureaucracies to command them. If that sounds like the unrealistic fantasy of people who yearn for a planet of equals, it's even more unrealistic to think that we can keep the status quo of domination, exploitation, and growth going without destroying our one and only planet. We need to institutionalize sharing and harmony instead of elite accumulation and exploitation, plain and simple.

In short, we are living in a new era that requires drastic change. It's time to talk honestly about the severity of the crisis we face and act with a real sense of urgency.

Thank you and I look forward to your questions.

[발표 5 | Speaker 5]

디지털전환 시대 불안정노동과 AI 알고리즘

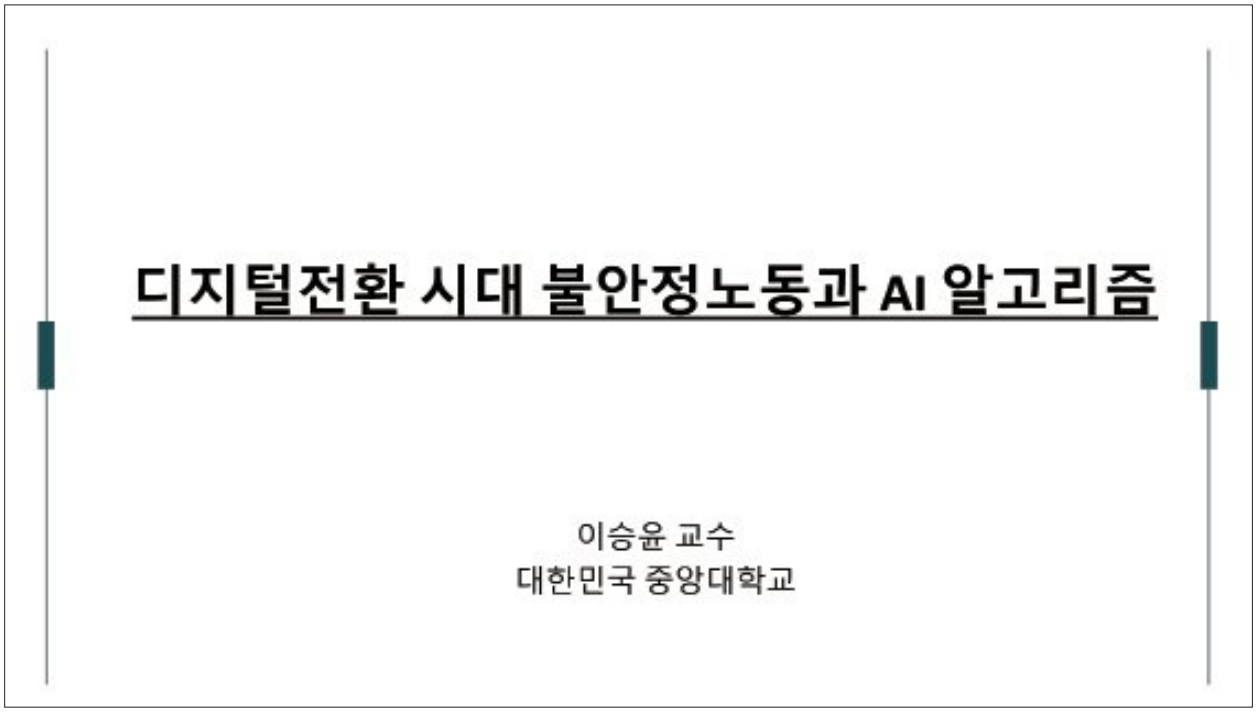
Precarious work and AI Algorithms in the Digital Transformation Era

이승윤
LEE Seung yoon

중앙대학교 사회복지학과 교수
Professor, Department of Social Welfare, Chung-Ang University

디지털전환 시대 불안정노동과 AI 알고리즘

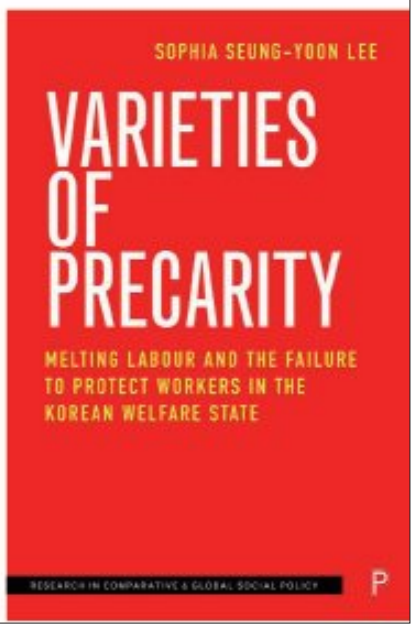
이승윤 | 중앙대학교 사회복지학과 교수



:변화하는 노동 형태

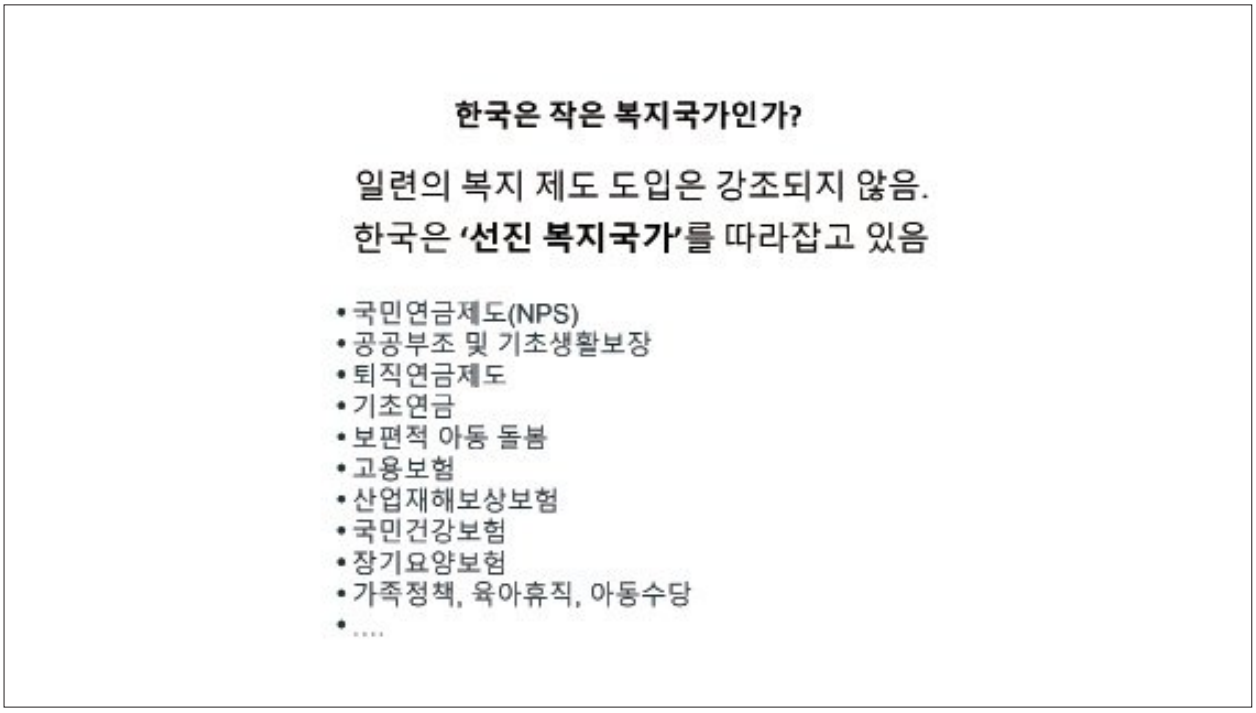
:전형적인 복지국가와의 <u>불일치</u>

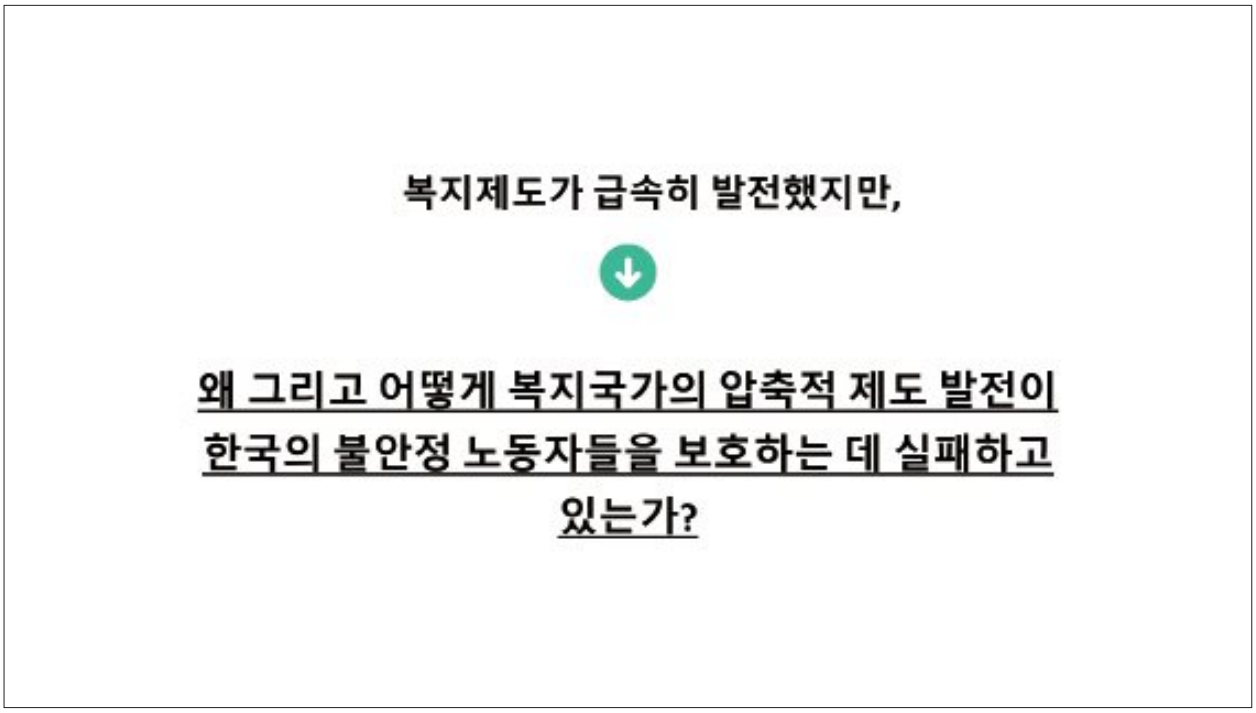
불안정"노동자"의 증가



한국의 압축적 복지국가 발전

복지 정치 없이, 한국은 약 30년간 압축적 복지국가 발전을 이룩함
(1960년대 사회보장법)


한국은 작은 복지국가인가?
일련의 복지 제도 도입은 강조되지 않음.
한국은 '선진 복지국가'를 따라잡고 있음
• 국민연금제도(NPS)
• 공공부조 및 기초생활보장
• 퇴직연금제도
• 기초연금
• 보편적 아동 돌봄
• 고용보험
• 산업재해보상보험
• 국민건강보험
• 장기요양보험
• 가족정책, 육아휴직, 아동수당
• ....


복지제도가 급속히 발전했지만,
왜 그리고 어떻게 복지국가의 압축적 제도 발전이 한국의 불안정 노동자들을 보호하는 데 실패하고 있는가?

액화노동

: 전통적인 노동 형태와 작업장을 둘러싼 다양한 경계의 해체

그림| 김상민 기자 일러스트(경향신문_2020년 신년특집 녹아내리는 노동 기획기사 시리즈)

새로운 노동 형태

이러한 표준고용관계(그리고 순수한 자영업)에서 벗어나는 다양한 노동 형태의 확산

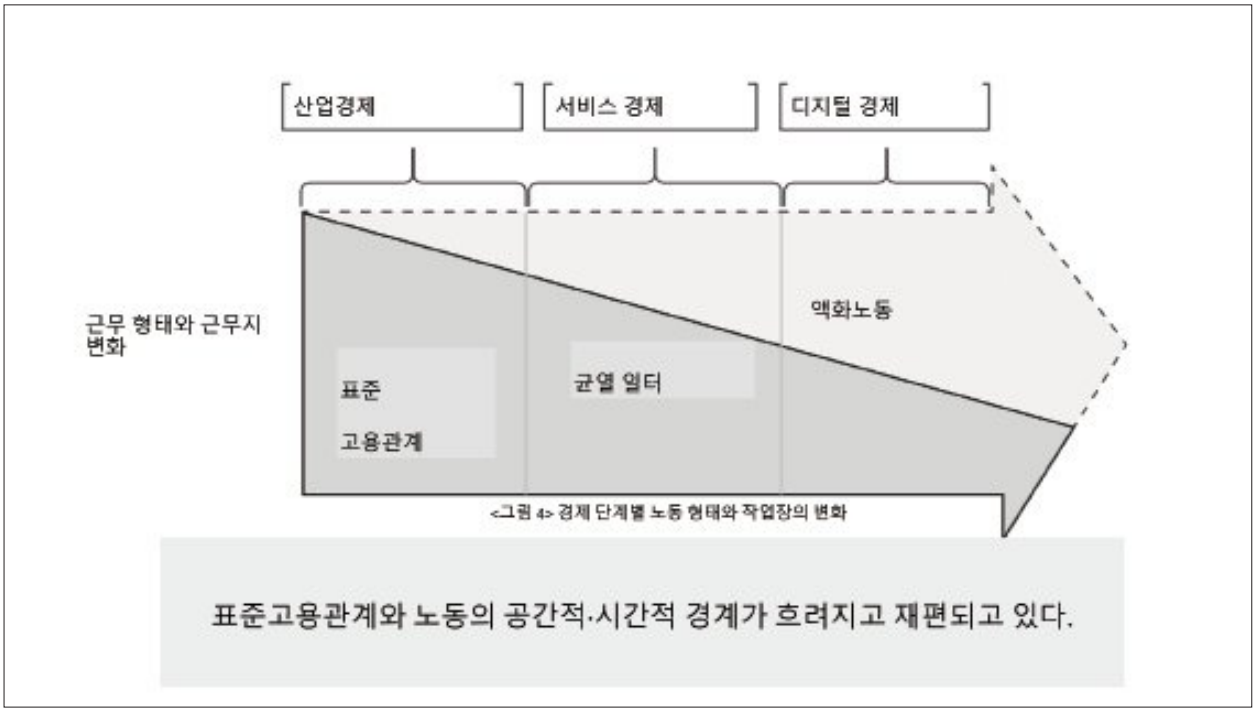
표준 노동과 직업 개념의 경계 흐려짐

- 근무시간과 휴식시간
- 고용과 실업
- 작업장과 개인공간, 공적 작업장
- 공식 노동과 비공식 노동
- 고객, 고용주, 피고용자

→

기존 제도의 배제

- 노동법제, 즉 최저임금, 근로시간 등
- 사회보장
- 교섭력의 약화
- 불안정성 증가



<그림 4> 경제 단계별 노동 형태와 작업장의 변화

표준고용관계와 노동의 공간적·시간적 경계가 흐려지고 재편되고 있다.

플랫폼 자본주의

2000년대 중반 이후 플랫폼 자본주의의 등장은 새로운 형태의 노동과 노동의 확장을 가져왔으며, 이러한 변화에 대한 정치경제학적 설명을 요구하고 있다.

산업 자본주의

- 전통적 산업 모델
- 이윤의 원천: 노동
- 임금기반 고용관계

- 데이터 주도 자본주의의 부상과 알고리즘 활용
- 이윤의 원천이 노동에서 데이터로 전환
- 데이터 포착과 통제를 위한 정교한 도구

플랫폼 자본주의

- AI 알고리즘 비즈니스 모델
- 이윤의 원천: 빅데이터
- 고용관계의 극도 분화

플랫폼 자본주의

› 작업이 아닌 과업 단위로 계약이 이루어지면서 노동자 정체성이 흐려지고, 노동자의 기술력과 교섭력이 약화되며, 소득이 불안정해짐

› 한국에서 노동의 탈상품화를 위한 제도적 보호는 여전히 (70-80년대에 제도화된) 표준고용관계에만 집중되어 있음

플랫폼 자본주의

› 알고리즘을 통한 보이지 않는 노동 통제와 미세 프로젝트, 미세업무의 확산, 크라우드 워크, 긱 워크의 증가.... *녹아내리는 노동*

이 새로운 AI 알고리즘 비즈니스 모델로부터 창출되는 이윤이 노동에 어느 정도 분배되고 있는가?

그림: 김상민 기자 일러스트(경향신문_2020년 신년특집 녹아내리는 노동 기획기사 시리즈)

AI 알고리즘이 노동자의 불안정성에 어떤 영향을 미치는가?

한국 심야배송 플랫폼 노동에서의 *알고리즘과 자기착취의 위험*

*기술 발전은 노동자에게 **자율성을 부여하는가?** (Rifkin, 1995)*
*아니면 오히려 더욱 미묘하고 정교한 방식으로 **자율성을 침식하는가?**(Srnicek, 2017; Lu, 2024)?*

이 연구는 디지털 기술이 제공하는 **유연성과 자율성의 약속**이 실제로는 **새로운 형태의 취약성과 착취**를 만들어내고 있는지를 살펴보는 것을 목표로 한다 (Droon, 2019).

유다영, 이승윤, 고태은 (중앙대학교) 2025 EASP/SPA 공동학회에서 발표한 논문

연구 배경: 한국의 심야배송 노동시장

왜 이것이 의미 있는 사례인가?

(1) 기술 발전을 통한 "자율성" 확대의 명확한 사례

- 주로 **직접적인 감독 없이** 야간에 업무 수행
- **스마트 기기와 디지털 플랫폼을 통해 독립적으로** 업무 관리

(2) 고용 유연성을 통한 노동 수요 충족

- 기업들은 증가하는 수요를 충족하기 위해 **긱 워커**에게 의존—**자율성의 약속**으로 포장됨

긱 워커를 홍보하는 데 사용된 채용 메시지

"내 집 근처에서 원하는 날 원하는 시간에"

(3) 자율성 약속 뒤에 숨겨진 착취의 근거

산업재해 건수

최근 5년 (2019~2023)

업무 관련 사망 **8 건**

⟹ *과로? 구조적 메커니즘!*

Q. 자율성이 어떤 방식으로 착취로 변질될 수 있는가?

⇒ 알고리즘으로 제조된 동의

제조된 동의 (Burawoy, 1979)

이 "게임"은 자율적인 것이 아니다 — 경영진에 의해 설계된 것이다.

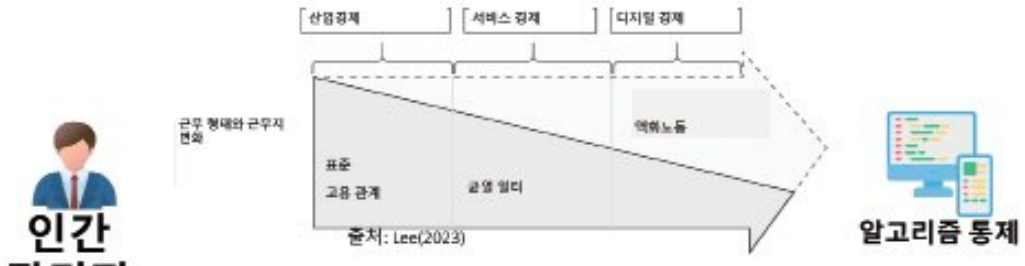
노동자들은 관리자에게 밀려나는 것이 아니다.

대신, 그들은 **게임의 룰**을 받아들이고 **경영진이 설계한** 시스템 내에서 승리하기 위해 노력함으로써 자발적으로 자기착취에 참여한다.

인간 관리자에서 알고리즘 통제로: 녹아 내리는 노동 시대의 일

정해진 근무시간도, 지정된 작업장도, 명확히 정의된 고용주도 없는 플랫폼 노동시장에서, 누가 동의를 제조했는가?

게임의 룰을 설계하고 시행하는 권한이 **인간 관리자에서 알고리즘으로** 이동했다.



"알고리즘이 만들어낸 동의"

디지털 노동의 맥락에서 자기착취는 어떻게 나타나는가?

플랫폼 노동에 대한 기존 연구를 바탕으로, 자기착취의 주요 패턴은 다음과 같다.

- **장시간 노동**
- **과도한 업무량**
- **휴식이나 회복시간 부족**
- **아픈 상태에서의 노동**

*이러한 행동들은 노동자들이 **알고리즘 통제와 디지털 노동 자율성** 하에서생산성에 대한 압박을 내재화하고 스스로를 너무 심하게 몰아붙이기 때문에 발생한다.*

알고리즘 상호작용이 디지털 플랫폼 노동에서 "자율적 노동"을 강화된 노동으로 변화시키는 방식

알고리즘이 만들어낸 동의

데이터 및 표본 설명

- 본 연구는 2024년 10월에 실시된 **한국의 심야 배달 플랫폼 노동자들을 대상으로 한 대규모 설문조사**를 바탕으로 하며, 총 **942개의 유효한 응답**을 수집했다.
- 본 설문조사는 주당 최소 하루 이상 일하고, 21:00~07:00 사이에 최소 한 시간 이상 배달 업무를 수행한 개인들을 대상으로 했다.
  → 표본이 **야간 근무 플랫폼 노동자들의** 핵심 집단을 반영하도록 보장한다.
- 설문지는 **94문항**으로 구성되었으며, 다음 내용을 포함한다.
  - 고용 유형 및 관계
  - 일반적인 근무 조건
  - 심리사회적 노동 환경
  - 노동자들의 인식과 경험

측정 및 조작화

매개변수: 알고리즘 통제

모든 문항은 4점 리커트 척도로 평가되었으며, 지수는 모든 문항의 평균 점수를 기반으로 한다.

범주		개념 (Fernández –Macías, 2023)	문항
방향	업무 할당	시간/교대 자동 할당	귀하의 근무 일정이나 시간이 앱이나 기기를 통해 자동으로 배정됩니까?
		활동 자동 할당	귀하의 배달 건수나 업무량이 기기를 통해 자동 배정됩니까?
	작업 과정	속도 자동 할당	귀하의 작업 속도나 물품 처리 속도가 기기나 앱의 영향을 받습니까?
		자동화된 지시	앱이나 기기에서 제공하는 자동화된 배달 경로나 지침을 따르십니까?
평가	업무 배정과 관련된 성과 평과	업무 할당에 사용되는 평점	앱 기반 성과 점수나 고객 평점이 향후 업무 배정에 사용됩니까?
		업무 취소에 사용되는 평점	최소 점수나 평점을 유지하지 못하면 앱에 의해 업무가 취소되거나 제한될 수 있습니까?

종속변수: 자기착취

각 하위 차원은 해당되면 1, 해당되지 않으면 0으로 코딩되었다. 모든 지표를 합산하여 0부터 4까지의 연속 점수를 생성했다.

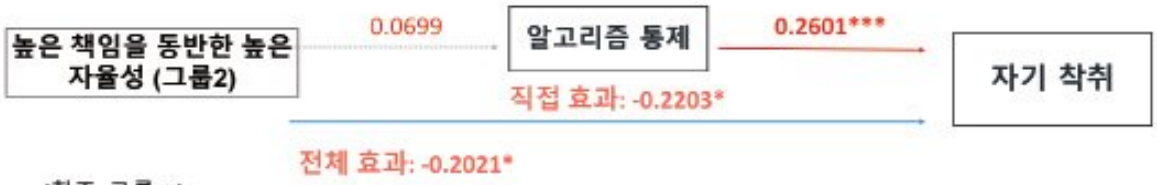
지표	질문	조작화
장시간 노동	주당 근무일 및 근무 시간	지난 주 심야 배달 근무 일수 총 시간(주간 & 야간): 주 52시간 이상 = 자기착취
과도한 업무량	일일 배달 건수	상위 25% = 자기착취
휴식 부족	일일 휴식시간	"휴식 없음" = 자기착취
아픈 상태에서 노동	아픈 상태에서 근무	"예" = 자기착취

결과

디지털 노동 자율성 유형의 **세 개의 뚜렷한 집단**

결과

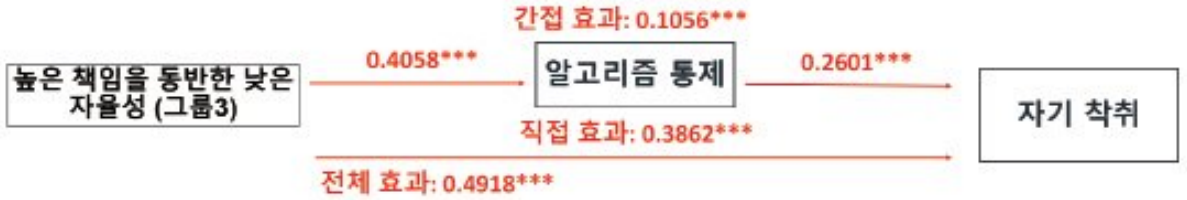
매개분석 결과



(참조. 그룹 1)

- 전반적으로, 높은 자율성은 자기착취를 감소시키는 구조적 효과를 가진다.
- 그러나 더 큰 알고리즘 통제 경향으로 인해 잠재적인 상쇄 효과가 있다—간접 경로는 통계적으로 유의하지 않았지만, 자기착취에 대한 알고리즘 통제의 강한 정적 효과가 자율성의 직접적 이익을 상쇄할 수 있다.

AI 알고리즘 매개분석 결과



(참조. 그룹 1)

- 두 가지 강화 경로
  - **직접 효과**: 낮은 자율성 + 높은 책임 → 자기착취
  - **간접효과**: 강력한 알고리즘 통제가 자기착취를 **증폭** 시킴
- 그룹 3은 대부분 **개인사업자**로 구성됨 — 가장 취약한 유형의 노동자들(Kim et al., 2022)

→ 이는 **"자기착취의 이중 구조"**를 형성함: 플랫폼 거버넌스가 구조적 및 알고리즘적 수단을 통해 취약성을 강화시킴

분석결과

알고리즘 매개분석 결과

효과 유형	그룹 2 높은 책임을 동반한 높은 자율성	그룹 3 높은 책임을 동반한 낮은 자율성
전체 효과	-0.2021* 기준집단 대비 전반적으로 낮은 자기착취	+0.4918*** 기준집단 대비 전반적으로 높은 자기착취
직접 효과	-0.2203*** 구조적 자율성이 자기착취를 직접적으로 감소시킴	+0.3862*** 구조적 의존성이 자기착취를 직접적으로 증가시킴
간접 효과	+0.0182 더 강한 알고리즘 통제를 통한 소폭 증가	+0.1056*** 훨씬 더 강한 알고리즘 통제를 통한 대폭 증가

분석결과

1. ***실질적 자율성이 중요하다.***
- 높은 책임을 동반한 높은 자율성 그룹의 노동자들은 더 적은 자기착취를 보임 (계수= –0.20)
- 업무, 시간, 의사결정에서의 자율성이 자기조절을 가능하게 함
- 알고리즘적 넛지는 존재하지만 자율성이 확보되면 제한적 영향만 미침

2. ***의존성이 취약성을 확대한다.***
- 높은 책임을 동반한 낮은 자율성 그룹은 현저히 더 많은 자기착취를 보임 (계수= 0.49)
- 실질적 자율성 부족 + 과중한 업무량과 비용 부담⇒ ***구조화된 의존성***
- 알고리즘 통제가 적시 업무 배정과 성과 연계 처벌을 통해 착취를 강화함 (그 효과가 총 영향의 ~20%를 설명)

3. ***불평등한 증폭기로서의 알고리즘 통제***
- **group**알고리즘 통제는 **낮은 자율성 그룹에서만** 자기착취를 유의미하게 매개함
- 높은 자율성 그룹에서는 유의미한 매개효과가 발견되지 않음
- 디지털 거버넌스 도구는 불안정성 하에서 동의를 증폭시킬 수 있음 → *"알고리즘이 제조한 동의"*

시사점

1. 사용자 규제와 복지국가/연대 책임으로부터의 자본 이탈

플랫폼 경제와 AI 알고리즘 비즈니스 모델은 고용주들이 이전이나 아웃소싱, 새로운 형태의 고용을 통해 복지 의무를 회피할 수 있게 한다.

› 고용주의 책임과 사회보장(연금, 건강보험, 실업급여)에 대한 기여를 의무화하는 **입법 및 규제 프레임워크**

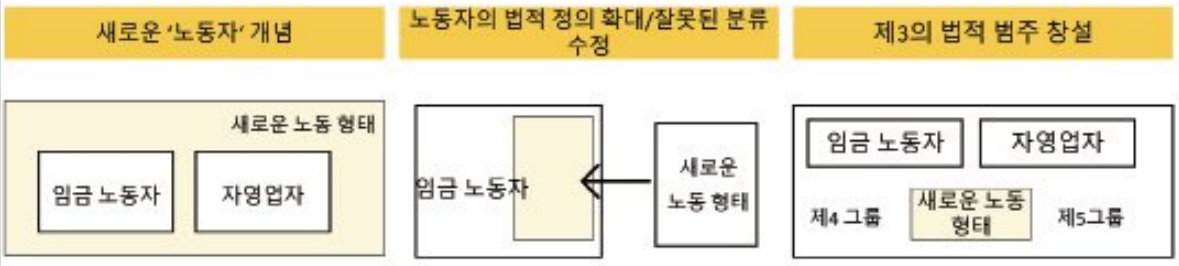
› 사회적 계약 회피를 막기 위한 특정 기여금과 같은 **재정 메커니즘**

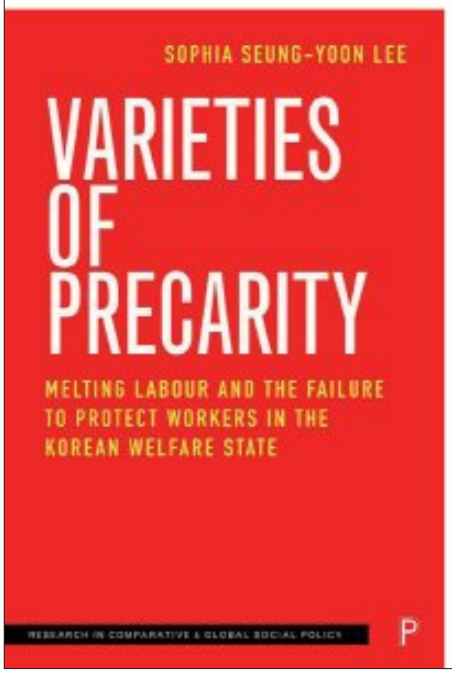
- **실질적 자율성 복원**
- 특히 **개인사업자**를 대상으로
- 일정 관리, 업무 선택, 의사결정에서의 자율성 보장

- **알고리즘 통제 규제**
- 알고리즘 의사결정의 **투명성** 요구
- 성과 지표에 **비례성 검증** 적용
- 비용과 위험을 노동자에게 전가하는 과도한 최적화 방지

2 연대의 토대로서의 보편적 사회보호

견고한 사회적 기반은 보편적이고 포용적인 보장에 달려있다: ***3가지 전략***



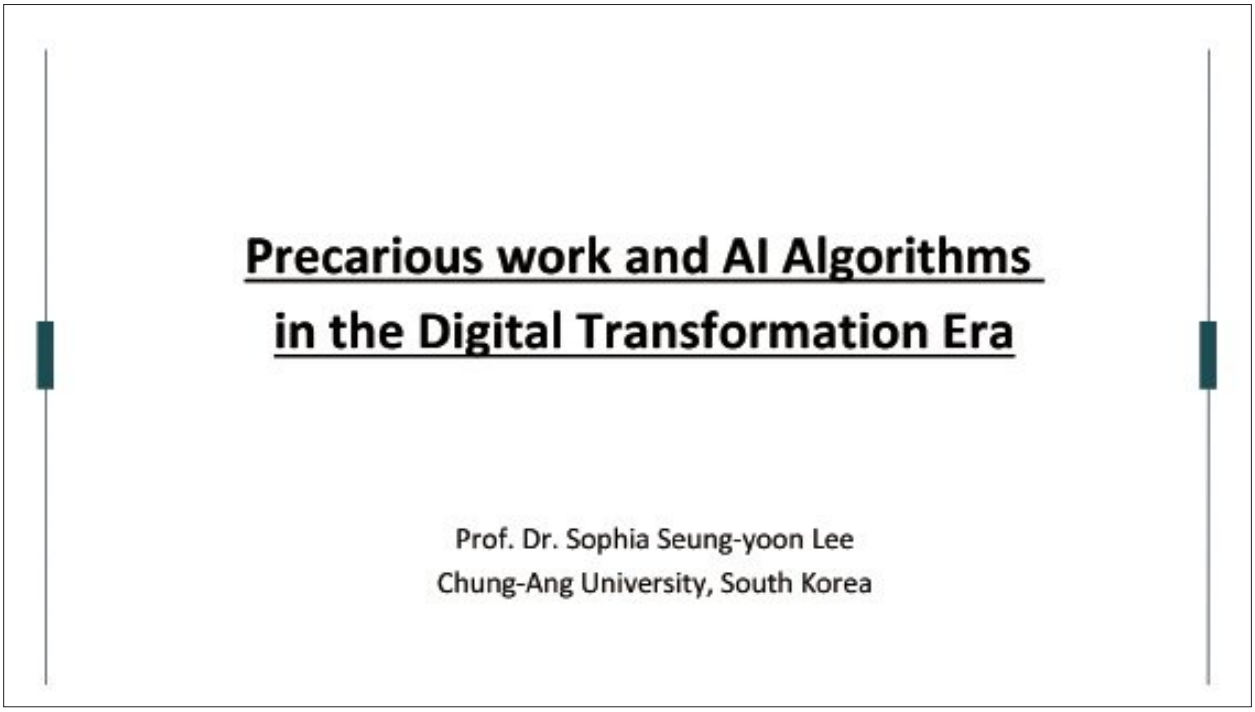


감사합니다.

leesophiasy@cau.ac.kr

Precarious work and AI Algorithms in the Digital Transformation Era

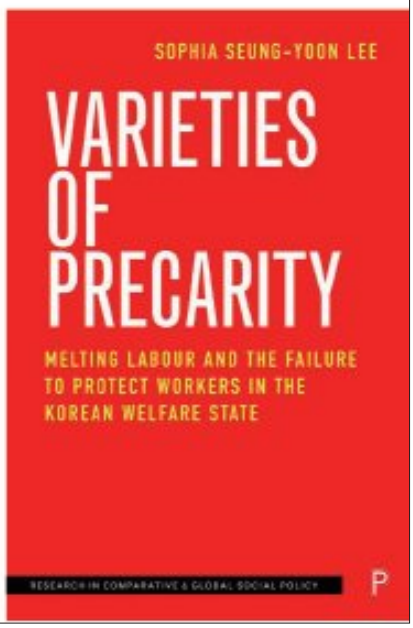
LEE Seung yoon | Professor, Department of Social Welfare, Chung-Ang University



:changing forms of work

:mismatch with the traditional welfare state

Increase of precarious workers



Compressed welfare state development in South Korea

Without the welfare politics, Korea achieved a compressed welfare state development in about 30 years (social security laws in1960s)

Korea as a small welfare state?

The introduction of a series of welfare institutions is often not emphasized.

Korea catching up with **'advanced welfare state'**

- National Pension Scheme (NPS)
- Public Assistance and Basic Livelihood Security
- Retirement Pension System
- Basic Pension
- Universal childcare
- Employment Insurance
- Industrial Accident Compensation Insurance
- National Health Insurance
- Long-Term Care Insurance
- Family policy, parental leave, child allowence,
- ....

Despite rapid development in the welfare institutions,

Why and how does the compressed institutional development of the welfare state fail to protect precarious workers in Korea?

Melting labour

: The dismantling of various boundaries surrounding the traditional forms of work and workplace

그림| 김상민 기자 일러스트(경향신문_2020년 신년특집 녹여내리는 노동 기획기사 시리즈)

New forms of WORK

the spread of various forms of work that deviate from these standard employment relationships (and also pure self-employment)

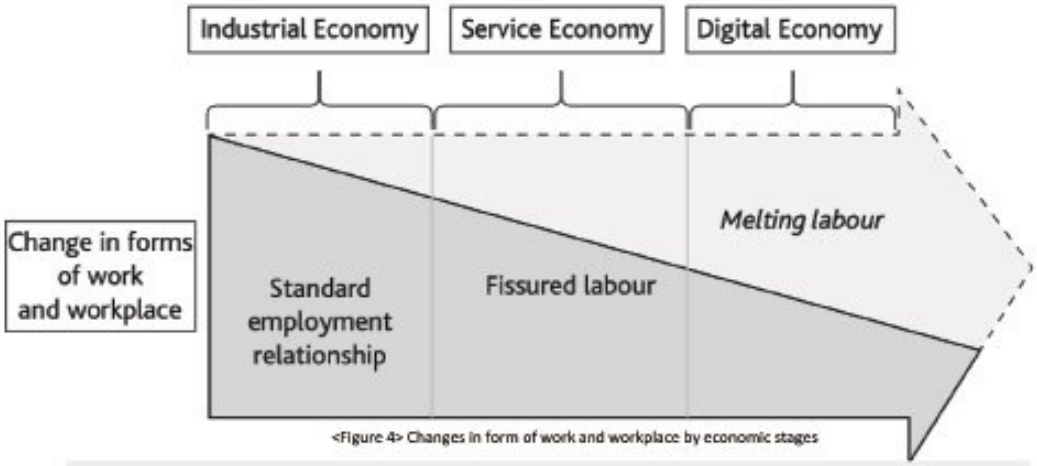
Blurrings around the concept of standard work and occupation

- Working time and resting time
- Employment and unemployment
- Work place and personal place, public workplace
- Formal work and informal work
- Clients, Employer, Employee

→

Exclusion of existing institutions

- labor regulations, ie. Minimum income, working time, etc.
- social protection
- Weakening of the bargaining power
- Rising precarity



<Figure 4> Changes in form of work and workplace by economic stages

the spatial and temporal boundaries of standard employment relations and work are being blurred and reshaped.

Platform Capitalism

Since the mid-2000s, the emergence of platform capitalism has brought about new forms and expansions of work, demanding a political-economic explanation of this transformation.

Industrial Capitalism		Platform Capitalism
› Traditional industrial model › Source of profit: labor › Wage-based employment relationship	› Rise of data-driven capitalism and use of algorithms › Shift from labor to data as profit source › Sophisticated tools for data capture and control	› AI Algorithm business model › Source of profit: big data › Hyper-fragmentation of employment relationships

Platform Capitalism

› Bluring of the workers identity, skill of a worker, weakening of bargaining power, and income instability as contracts are made in units of tasks rather than jobs

› The institutional protection for the de-commodification of work is still concentrated on the standard employment relationship in Korea (institutionalized in the 70s and 80s)

Platform Capitalism

› Invisible labor control through algorithms and the expansion of micro projects, micro tasks, increase of crowd works, gig works....*melting labour*

To what extent are the profits derived from this new AI Algorithm bussiness model is being distributed to labour?

그림: 김상민 기자 일러스트(경향신문_2020년 신년특집 녹아내리는 노동 기획기사 시리즈)

How AI Algorithm effects workers' precarity?

Algorithm and risk of self-exploitation in Korean Midnight Delivery Platform Labor

*Does technological advancement **grant autonomy** to workers (Rifkin, 1995)?*
*Or rather, **erode autonomy** in more subtle and sophisticated ways?(Srnicek, 2017; Lu, 2024)?*

This study aims to examine whether the **promise of flexibility and autonomy** offered by digital technologies is, in practice, producing **new forms of vulnerability and exploitation** (Droon, 2019).

Yu, Dayoung., Lee, Seung-yoon., Ko, Taieun (Chung-Ang University) Paper presented at the 2025 EASP/SPA joint conference

Research Context: South Korea's Midnight Delivery Labor Market

Why is this a meaningful case?

(1) A clear case of expanding "autonomy" through technological advancement

- Work is performed at night, usually **without direct supervision**
- Tasks are managed **independently via smart devices & digital platform**

(2) Labor demand is met through employment flexibility

- Firms turn to **gig workers** to meet growing demand—framed by the **promise of autonomy.**

Recruitment messages used to promote gig work

"Work whenever you want, near your home"

(3) Evidence of exploitation behind the promise of autonomy

Number of industrial accidents

Over the past 5 years (2019~2023)

Work-related deaths **8 Cases**

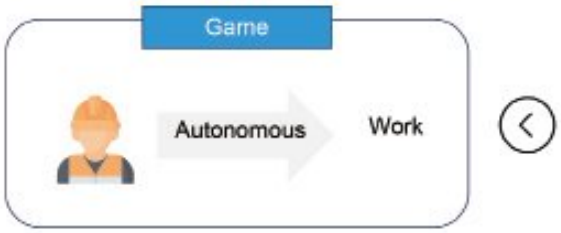
⟹ *Overwork? Structural mechanism!*

Q. In what ways can Autonomy turn into Exploitation?"

⇒ *Algorithm Manufactured Consent*

Manufactured Consent (Burawoy, 1979)

The "game" isn't free — it's designed by management.



Structured by

Human Managers

Workers are not pushed by managers.

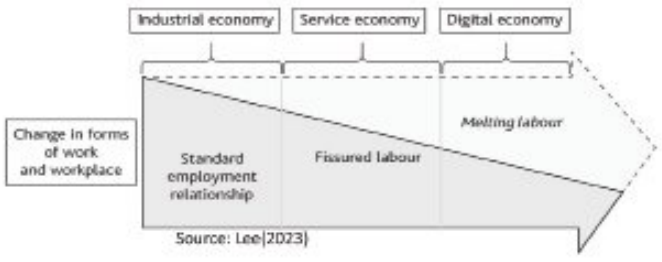
Instead, they voluntarily engage in self-exploitation
by accepting **the rules of the game** and striving to win within a system **designed by management**.

From Human Managers to Algorithmic control: Work in the Melting Labor Era

In a platform labor market where there are no fixed working hours, no designated workplace, and no clearly defined employer, who manufactured consent?

The authority to design and enforce the rules of the game has shifted: **from human managers to algorithms.**

Human managers



Algorithmic Control

"Algorithm Manufactured Consent"

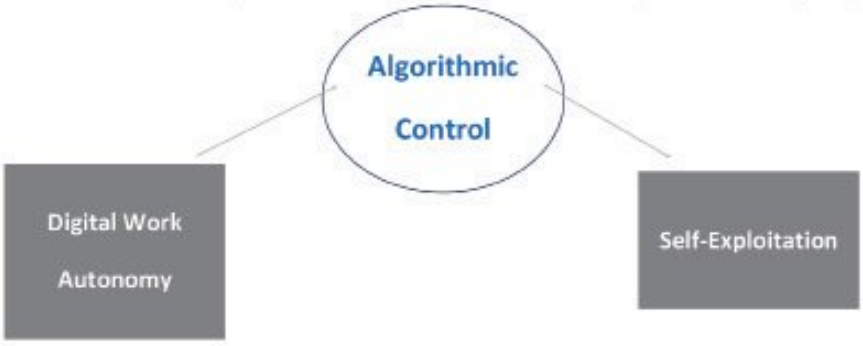
How does self-exploitation manifest in the context of digital labor?

Drawing on prior research on platform work, key patterns of self-exploitation include:

- **Long working hours**
- **Excessive workloads**
- **Lack of rest or recovery time**
- **Working while sick**

These behaviors happen because workers,
*under **algorithm control** and **digital work autonomy**,*
Internalize the pressure to be productive and push themselves too hard.

How algorithmic interaction transforms "autonomous labor" into intensified labor in digital platform work:



Algorithm Manufactured Consent

Data and Sample Description

- This study draws on a large-scale survey of **midnight delivery platform workers in South Korea**, conducted in October 2024, with a total of **942 valid responses**.
- The survey targeted individuals who work at least one day per week, and who performed delivery tasks for at least one hour between 21:00~07:00
  → This ensures the sample reflects the core group of **night-shift platform laborers**.
- The questionnaire consisted of **94 items**, covering:
  - Employment type and relationship
  - General working conditions
  - Psychosocial labor environment
  - Workers' perceptions and experiences

Measurement and Operationalization

Mediator Variable : Algorithmic Control

All items were rated on a 4-point Likert scale, with the index based on the average score of all items.

Category		Concept (Fernández–Macías, 2023)	Questions
Direction	Task Allocation	Automatic allocation of time/shifts	Is your work schedule or hours automatically assigned via an app or device?
		Automatic allocation of activities	Is your number of deliveries or workload automatically assigned through a device?
	Work Process	Automatic allocation of speed	Is your work speed or item handling pace influenced by a device or app?
		Automated direction	Do you follow automated delivery routes or instructions from an app or device?
Evaluation	Performance Evaluation linked to task assignment	Rating used to allocated work	Are app-based performance scores or customer ratings used to assign your future work?
		Rationg used to cancel work	If you don't maintain a minimum score or rating, can your tasks be canceled or restricted by the app?

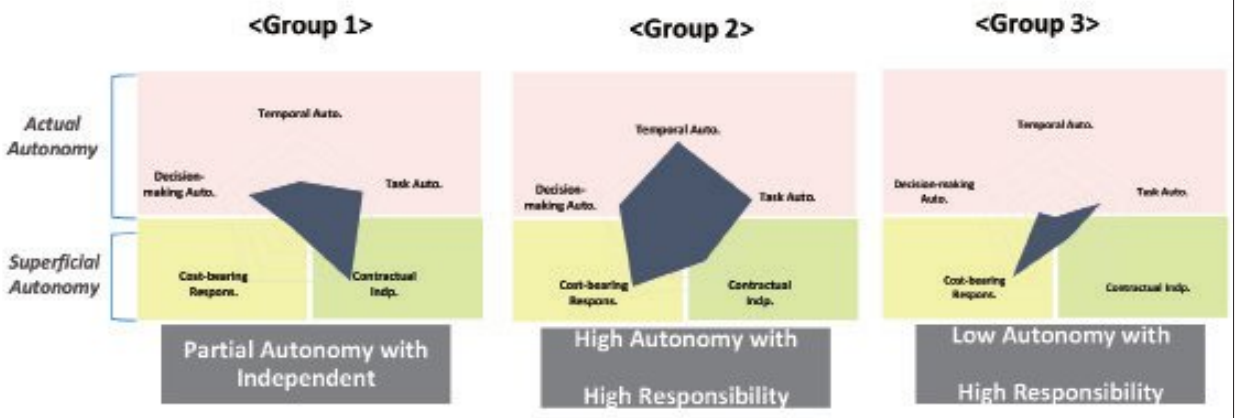
Dependent Variable
: Self Exploitation

Each sub-dimension was coded as 1 if applicable and 0 if not.
All indicators were summed to create a continuous score from 0 to 4.

Indicators	Questions	Operations
Long working hours	Working days per week & working hours	Number of days worked in midnight delivery during the past week Total hours(day & night); 52+ hours/week = Self-exploitation
Excessive work loads	Daily delivery count	Top 25% = Self-exploitation
Lack of rest	Daily Rest time	"No rest" = Self-exploitation
Working while sick	Working while Sick	"YES" = Self-exploitation

Result

three distinct groups of digital work autonomy types



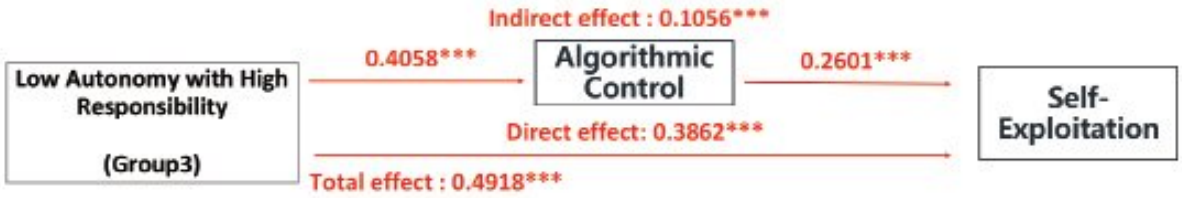
Result

Mediation Analysis Result

(Ref. Group1)

- Overall, higher autonomy has a structural effect in reducing self-exploitation.
- However, due to a tendency toward greater algorithmic control, there is a potential offsetting effect—although the indirect path was not statistically significant, the strong positive effect of algorithmic control on self-exploitation may counteract the direct benefit of autonomy.

AI Algorithm Mediation Analysis Result



(Ref. Group1)

- Two reinforcing pathways:
  - **Direct effect**: Low autonomy + high responsibility → self-exploitation
  - **Indirect effect**: Strong algorithmic control **amplifies** self-exploitation
- Group 3 consists mostly of **independent contractors** — the most vulnerable type of workers (Kim et al., 2022)

→ This forms a **"dual structure of self-exploitation"**: Platform governance intensifies vulnerability through structural and algorithmic means

Findings

Algorithm Mediation Analysis Result

Effect Type	Group 2 High Autonomy with High Responsibility	Group 3 Low Autonomy with High Responsibility
Total effect	-0.2021* overall lower self-exploitation than reference	+0.4918*** overall higher self-exploitation than reference
Direct effect	-0.2203*** structural autonomy directly reduces self-exploitation	+0.3862*** structural dependency directly raises self-exploitation
Indirect effect	+0.0182 small increase via stronger algorithmic control	+0.1056*** large increase via much stronger algorithmic control

Findings

1. *Actual Autonomy Matters*
- Workers in the High Autonomy with High Responsibility group show less self-exploitation (Coeff= –0.20)
- Autonomy in task, time, and decision-making enables self-regulation
- Algorithmic nudges are present but have limited impact when autonomy is secured

2. *Dependency Magnifies Vulnerability*
- Low autonomy with High Responsibility group shows significantly more self-exploitation (Coeff= 0.49)
- Lack of actual autonomy + heavy workload and cost burden ⟹ ***Structured dependency***
- Algorithmic control intensifies exploitation through just-in-time tasking and performance-linked penalties (and Its effect explains ~20% of total impact)

3. *Algorithmic Control as an Unequal Amplifier*
- Algorithmic control significantly mediates self-exploitation **only in the Low Autonomy group**
- No significant mediation was found in the high autonomy group
- Digital governance tools can amplify consent under precarity → *"Algorithm-manufactured consent"*

Implications

1. Constraining Employers and Capital Exit from Welfare State/Solidarity Responsibilities

Platform economy, AI Algorithm business model allow employers to evade welfare obligations by relocating or outsourcing, hiring as new forms for work,

- **Legislative and regulatory frameworks** that mandate employer's responsibilities and contributions to social security (pensions, health, unemployment)
- **Fiscal mechanisms** like earmarked contributions to block social contracting out

- Restore Actual autonomy

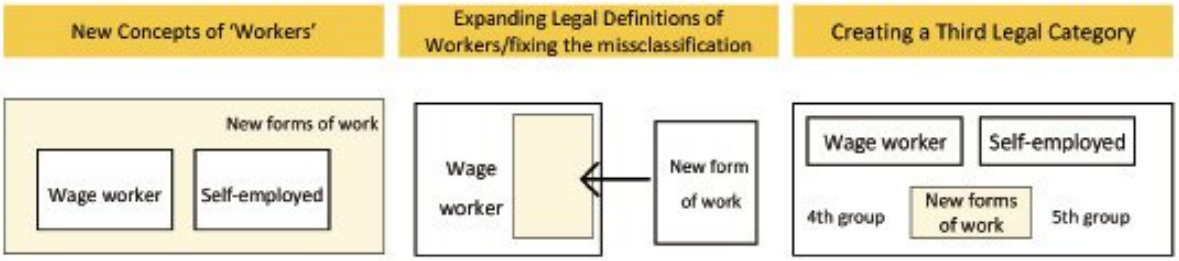
- Especially for **independent contractors**
- Ensure autonomy in scheduling, task choice, and decision-making

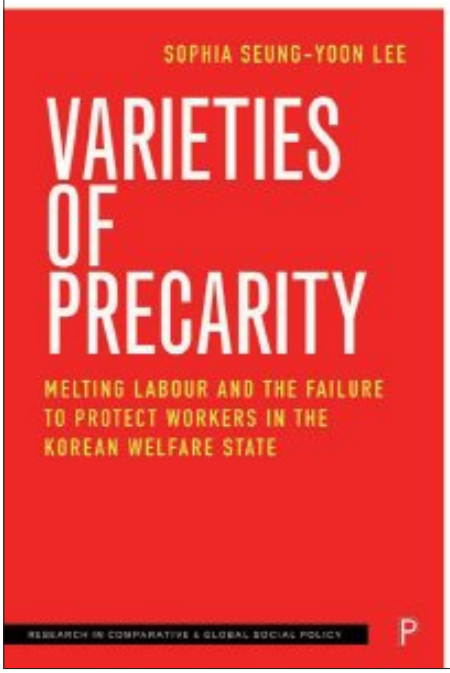
- Regulate Algorithmic control

- Require **transparency** in algorithmic decision-making
- Apply **proportionality tests** to performance metrics
- Prevent over-optimization that shifts cost and risk onto workers

2 Universal Social Protection as Foundation for solidarity

A robust social fabric depends on universal, inclusive coverage: ***3 strategies***





Thank you

leesophiasy@cau.ac.kr

세션 3
Session 3

기술발전과 미래 그리고 대응
Responding to Human Rights Challenges in the Digital Age

사회자 Moderator

안성율 | 국가인권위원회 정책교육국장
AN Seong Ryul | Director General, Policy Bureau, NHRCK

발표자 Speaker

장여경 | 정보인권연구소 상임이사
CHANG Yeo-kyung | Executive Director, Institute for Digital Rights

유승익 | 명지대학교 법학과 객원교수
YOO Seung-ik | Guest Professor, Department of Law, Myongji University

넬레 루켄스 | 유럽 국가인권기구 연합 AI와 인권 실무그룹 의장
Nele ROEKENS | Chair of the Working Group on AI, ENNHRI

세션 3 | Session 3

[사회자_Moderator]

안성율

AN Seong Ryul

국가인권위원회 정책교육국장

Director General, Policy Bureau, NHRCK

[주요경력]

사법연수원 제32기 수료, 변호사

국가인권위원회 인권정책과장, 행정법무담당관, 운영지원과장

국가인권위원회 침해조사국장

[Career]

Seung Ryul AN completed the 32nd Judicial Research and Training Institute and is a licensed Attorney-at-Law.

He has served as Director of the Human Rights Policy Division, Administrative and Legal Affairs Officer, Director of General Affairs, and Director General of the Civil and Political Rights Bureau at the National Human Rights Commission of Korea (NHRCK).

He is currently the Director General of the Human Rights Policy Bureau, NHRCK.

세션 3 | Session 3

[발표자_Speaker]

장여경

CHANG Yeo-kyung

정보인권연구소 상임이사

Executive Director, Institute for Digital Rights

[주요경력]

장여경은 교육공학(Educational Technology)과 과학기술학(Science, Technology, and Society)을 공부했다. 2018년 정보인권연구소에서 상근을 시작하였고, 그전에는 1996년 정보인권단체 진보네트워크센터(Jinbonet)의 창립에 참가한 이래 줄곧 활동가로 상근했었다.

한국 사회에 정보인권(Digital Rights)의 개념을 소개하고 정책에 반영하기 위해 노력해 왔다. 서울특별시교육청에서 학생인권위원회 비상임위원을, 개인정보보호위원회에서 비상임위원을 맡기도 하였다. 현재는 국가인권위원회 정보인권전문위원회의 위원이다.

[Career]

Yeo-kyungChangstudied Educational Technology and Science, Technology, and Society. She began working full-time at the Information Human Rights Research Institute in 2018, and prior to that, she had been working as a full-time activist since participating in the founding of the progressive network center Jinbonet, a digital rights organization, in 1996. She has been working to introduce the concept of Digital Rights to Korean society and reflect it in policy. She served as a non-standing member of the Student Human Rights Committee at the Seoul Metropolitan Office of Education and as a non-standing member of the Personal Information Protection Commission. She is currently a member of the Digital Rights Expert Committee at the National Human Rights Commission of Korea.

세션 3 | Session 3

[발표자_Speaker]

유승익

YOO Seung-ik

명지대학교 법학과 객원교수

Guest Professor, Department of Law, Myongji University

[주요경력]

유승익은 2011년, 고려대학교에서 헌법학을 전공했다. 그는 2012년부터 신경대학교(현 화성의과학대학교) 법학과 및 경찰행정학과 조교수를 역임했고, 2021년부터 2025년까지 한동대학교 BK21 연구교수를 지냈고, 현재 명지대학교에서 헌법과 인권 관련 강의와 연구를 계속하고 있다. 참여연대 사법감시센터 소장직을 맡고 있다.

[Career]

Seung-ik Yoo majored in constitutional law at Korea University in 2011. He served as an assistant professor in the Department of Law and Police Administration at Shingyeong University (now Hwaseong University of Medical Sciences and Technology) from 2012, worked as a BK21 research professor at Handong Global University from 2021 to 2025, and is currently continuing his lectures and research on constitutional law and human rights at Myongji University. He serves as the director of the Judicial Watch Center at People's Solidarity for Participatory Democracy (PSPD).

세션 3 | Session 3

[발표자_Speaker]

넬레 루켄스
Nele ROEKENS

유럽 국가인권기구 연합 AI와 인권 실무그룹 의장
Chair of the Working Group on AI, ENNHRI

[주요경력]

넬레(Nele)는 유럽국가인권기구 네트워크(ENNHRI)의 AI 워킹그룹 의장을 맡고 있다. 이 직책을 통해 그녀는 AI와 신기술 분야에서 40여 개 독립 공공기관들의 공동 노력을 총괄하고 있다. 또한 유럽평의회 인공지능위원회에서 ENNHRI 대표로도 활동하고 있다.
자국에서는 벨기에 평등기구이자 국가인권기구인 유니아(Unia)의 인공지능 팀장을 맡고 있다.
넬레는 유럽평의회 인공지능, 평등, 차별 전문가위원회(GEC/ADI-AI)의 독립 전문가로 임명되었으며, UN 사무총장 산하 인공지능 고위급 자문기구의 전문가 네트워크 멤버로도 활동했다. 이러한 역할을 통해 그녀는 AI와 신기술의 개발, 배치, 규제에서 인권 기반 접근법을 옹호하고 있다.
[현재 역할 이전에 그녀는 UN 인권이사회, CNN, 유로저스트(Eurojust), 그리고 토고의 불처벌 척결을 위한 NGO 연합체인 CACIT 등 저명한 기관들에서 경험을 쌓았다.]

[Career]

Nele is chair of the Working Group on AI of the European Network of National Human Rights Institutions' (ENNHRI). In this role she guides the collaborative efforts of over 40independent public institutions in matters on AI and emerging technologies. This role extends to representing ENNHRI at the Council of Europe's Committee on Artificial Intelligence.
At the national level, Nele leads the Artificial Intelligence team at Unia, the Belgian Equality Body and National Human Rights Institution.
Nele is appointed as an independent expert for the Council of Europe's Committee of Experts on Artificial Intelligence, Equality, and Discrimination (GEC/ADI-AI) and was member of a network of experts for UN Secretary-General's High-Level Advisory Body on Artificial Intelligence. Through these roles, Nele advocates for a human rights based approach in the development, deployment and regulation of AI and emerging technologies.
[Prior to her current roles, she gained experience at prominent institutions including the United Nations Human Rights Council, CNN, Eurojust, and CACIT, a collective of NGOs that combat impunity in Togo.]

[발표 1 | Speaker 1]

AI 인권거버넌스:
인권기반 접근과 영향받는 사람

AI Human Rights Governance: Human Rights-Based Approach and Affected Persons

장여경
CHANG Yeo-kyung

정보인권연구소 상임이사
Executive Director, Institute for Digital Rights

AI 인권거버넌스: 인권기반 접근과 영향받는 사람

장여경 | 정보인권연구소 상임이사

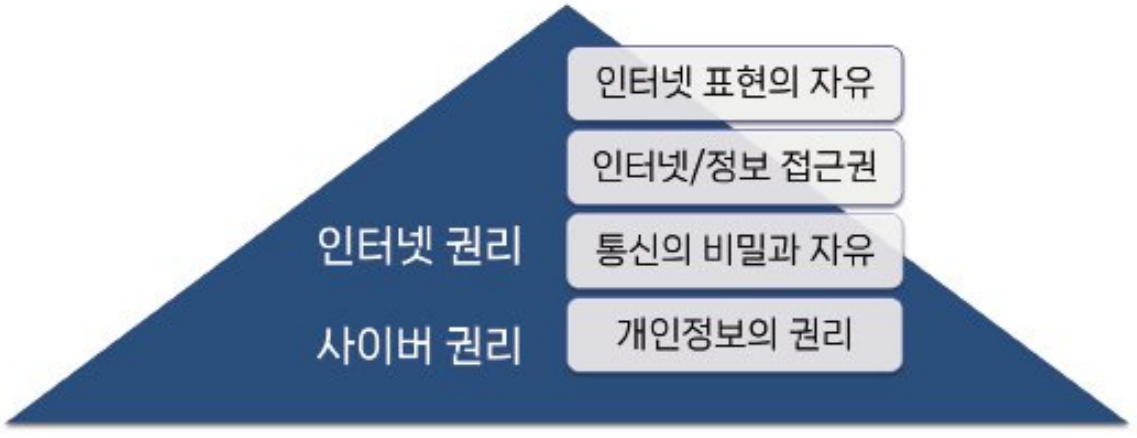
정보인권의 발전

The Evolution of Digital Rights

2

권리주체의 발견과 권리 식별의 과정

"인터넷 이용자"의 권리로 시작된 요구



3

디지털 전환

- 이전까지 디지털 기술과 관련이 적었던 인간의 활동과 사물, 사회 전체가 네트워크로 연결되고, 노동부터 여가까지 삶의 모든 영역이 디지털 네트워크 기반 위에서 이루어짐

- 인터넷 이용자의 권리로부터 모든 시민의 보편적인 디지털 권리 보장에 대한 요구로 확대

4

국가인권기구의 정보인권 식별

국가인권위원회 (정보인권보고서, 2013)

"정보인권은 정보통신기술에 의하여 디지털화된 정보가 수집, 가공, 유통, 활용되는 과정과 그 결과로 얻어진 정보가치에 따라 인간의 존엄성이 훼손되지 않고, 자유롭고 차별없이 이용할 수 있는 기본적 권리이다."

5

정보인권의 명시 요구

대통령 헌법개정안 (2018)

- 모든 국민은 알권리를 가진다.
- 모든 사람은 자신에 관한 정보를 보호받고 그 처리에 관하여 통제할 권리를 가진다.
- 국가는 정보의 독점과 격차로 인한 폐해를 예방하고 시정하기 위하여 노력해야 한다.
- 언론 · 출판 등 표현의 자유는 보장되며, 이에 대한 허가나 검열은 금지된다.

6

AI 시대와 인권 영향

유엔 극빈과 인권에 관한 특별보고관 보고서 (2019)

- 첫째, 일반 인구 집단의 행동에서 도출된 예측을 바탕으로 개인의 권리를 결정함으로써 많은 문제가 야기되고 있다.
- 둘째, 기술의 기능과 특정 점수 또는 분류에 도달하는 방법은 종종 비밀로 지정되어 있어, 정부와 민간 행위자의 권리 침해 가능성을 파악하기 어렵다.
- 셋째, 위험점수를 계산하고 수요를 분류하는 것이 기존의 불평등과 차별을 강화하거나 악화시킬 수 있다.

7

AI에 대한 인권기반 접근

HumanRights-Based Approach to AI

8

윤리 기반 접근

빅테크와 정부 주도

기업

`아실로마 AI 원칙` 1년 뒤…카카오 `알고리즘 윤리 규범` 발표

입력 : 2018-02-02 13:20:49 수정 : 2018-02-02 13:33:51

가

과학기술정보통신부

보도자료

디지털뉴딜

보도일시	2020. 12. 23.(수) 11:00 이후 부터 보도해 주시기 바랍니다.		
배포일시	2020. 12. 22.(화) 14:00	담당부서	인공지능기반정책과
담당과장	김경만(044-202-6270)	담 당 자	유주연 사무관(044-202-6276)

과기정통부, 사람이 중심이 되는
「인공지능(AI) 윤리기준」 마련

-공개 공청회(12.7) 등 각계 전문가·시민 공개 의견수렴 거쳐 발표-

'인간성(Humanity)을 위한 인공지능(AI)'의 3대 원칙·10대 요건 담아

9

'자율적' 준수

과학기술정보통신부 (인공지능 윤리기준, 2020)

ㅇ **(목표 및 지향점)** ① 모든 사회 구성원이 ② 모든 분야에서 ③ 자율적으로 준수하며 ④ 지속 발전하는 윤리기준을 지향한다.

- [3]구속력 있는 '법'이나 '지침'이 아닌 **도덕적 규범이자 자율규범**으로, **기업 자율성을 존중**하고 인공지능 **기술발전을 장려**하며 **기술과 사회변화에 유연하게 대처**할 수 있는 윤리 담론을 형성

10

국제인권기구의 '자율적 윤리' 비판

유엔 의사표현의자유 특별보고관 보고서 (2018)

- "민간과 공공 부문이 '윤리' 규범을 내세우려는 배경에 인권기반 규제에 대한 저항이 있다."

유엔 극빈과 인권에 관한 특별보고관 보고서 (2019)

- "빅테크가 인권 의무로부터 거의 면제되어 있다(almost human rights free-zone)."

11

AI 의사결정에 대한 권리구제 요구

사례) 美 휴스턴 교육청 EVAAS 사건

- 휴스턴 연방지방법원 (2017)
  민간기업에서 조달한 교사평가 알고리즘으로
  고용 관련 공공 의사결정 내릴 때
  "헌법상 적법절차 준수해야"

➔ 중요한 공공 의사결정에서 비밀 알고리즘 중단

12

인권기반 접근 제안

유엔 의사표현의자유 특별보고관 보고서 (2018)

- 머신러닝의 적응성으로 인해, 사람이 AI 시스템의 목표와 결과를 정의하는 과정에서 점차 배제될 수 있다 ... 이는 AI의 **투명성, 책무성** 및 효과적인 **구제**수단에 대한 접근을 어렵게 만들 수 있다
- **국가와 기업**이 **인권영향평가** 등 인권법적 **의무**를 준수해야 한다

13

인권기반대응 제안

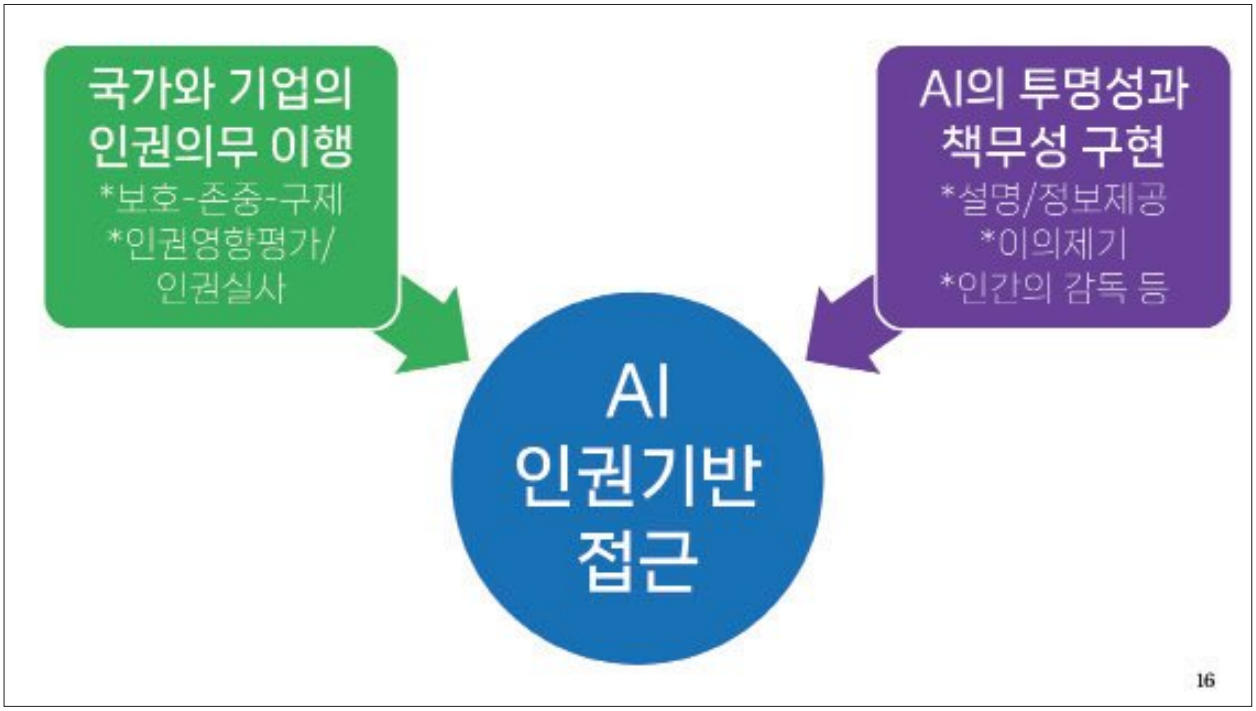
유엔 사무총장 보고서 (2020)

- 사람을 **권리주체** 개인으로 대우하고, 역량을 강화하며,
- 권리를 행사할 수 있는 **법적·제도적** 환경을 조성하여
- 인권침해를 **구제**하는 접근법

14

AI 인권규범의 발전

국가인권위원회 (2022) <인공지능 개발과 활용에 관한 인권 가이드라인>	유럽평의회 (2024) <인공지능과 인권·민주주의·법치주의에 관한 기본협약>
1. 인간의 존엄성 존중 2. 투명성과 설명 의무 3. 자기결정권의 보장 4. 차별 금지 5. 인공지능 인권영향평가 시행 6. 위험도 등급 및 관련 법제도 마련	1. 인간 존엄성과 개인 자율성 2. 투명성 및 감독 3. 책무성 및 책임성 4. 평등 및 차별 금지 5. 사생활 존중 및 개인정보보호 6. 신뢰성 7. 안전한 혁신


국가와 기업의
인권의무 이행
*보호-존중-구제
*인권영향평가/
인권실사
AI의 투명성과
책무성 구현
*설명/정보제공
*이의제기
*인간의 감독 등
AI
인권기반
접근
16


AI 시대의
인권 거버넌스
Human Rights Governance
in the Age of AI
17

자동화된 결정의 영향

과학기술정보통신부 www.msit.go.kr

보도자료

보도일시	2020. 4. 14.(화) 조간(온라인 4. 13. 12:00)부터 보도해 주시기 바랍니다.		
배포일시	2020. 4. 13.(월) 09:00	담당부서	디지털사회기획과
담당과장	박민하(044-202-6130)	담 당 자	박 솔 사무관 (044-202-6135)

머신러닝으로 부정수급 탐지, AI로 강력범죄 전자감독…
과기정통부, 공공서비스 혁신에 207억원 투입

- ICT기반 공공서비스 촉진사업 온라인 사업설명회(4.16) 개최 -

법제처 **보도자료** 보다나은 정부 한국판뉴딜

배포 시점부터 보도해 주시기 바랍니다.

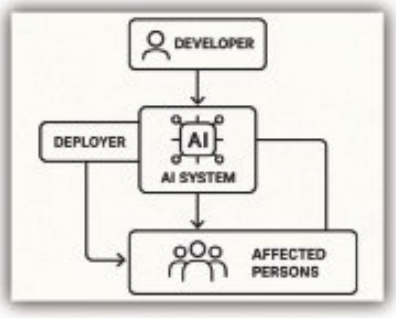
제20조(자동적 처분) 행정청은 법률로 정하는 바에 따라 완전히 자동화된 시스템(인공지능 기술을 적용한 시스템을 포함한다)으로 처분을 할 수 있다. 다만, 처분에 재량이 있는 경우는 그러하지 아니하다.

3월 23일 「행정기본법」 공포 · 시행

18

AI 시대 권리주체 식별

o **"영향받는 자"란** 인공지능제품 또는 인공지능서비스에 의하여 자신의 생명, 신체의 안전 및 기본권에 중대한 영향을 받는 자를 말한다(**한국 인공지능기본법** 제2조).



19

영향받는 사람의 참여

유엔 사무총장 (2020)

"국가는 권리주체, 특히 가장 큰 영향을 받거나 부정적인 결과를 겪을 가능성이 높은 **권리주체가** 개발 과정에 효과적으로 **참여하고 기여**할 수 있는 기회를 창출하고 특정한 신기술의 채택을 촉진해야 한다. **국가는 참여 보장과 포용적 의견수렴**을 통해서, 경제적 효율성, 환경적 지속 가능성, 포용성 및 형평성을 갖춘 균형적이고 통합적인 지속 가능 개발 목표에 있어 어떤 기술이 가장 적절하고 효과적인지 결정할 수 있다."

20

EU 사례) 영향받는 사람에게 공개/협의

직장 고위험AI의 노동자 공개

직장에서 고위험 AI시스템의 서비스를 제공하거나 사용하는 회사는 그 전에 노동자 대표자와 대상 노동자에게 고위험 AI시스템 사용의 대상이 될 것이라는 사실을 알려야 한다.

영향받는 사람의 참여와 협의 절차로서 기본권영향평가

인공지능 인권 위험을 완화하는 조치 과정에 영향받는 사람을 비롯한 이해관계자를 참여시키고 의견을 수렴한다.

21

영향받는 사람의 구제

유럽연합 인공지능법 ~ 유럽평의회 AI 기본협약

- 인권에 중대한 영향을 미칠 가능성이 있는 AI 시스템 관련 **정보를 관할 당국에 제공**하고, 특정한 경우 영향 받는 사람에게도 제공한다.
- 인공지능에 기반한 **결정으로부터 영향 받는 사람**은 결정에 대하여 **이의를 제기**할 수 있고, 특정한 경우 시스템 자체에 대한 **정보를 제공**받을 수 있다.
- 관련된 사람은 관할 당국에 **진정을 제기**할 수 있다.

22

AI 인권 거버넌스의 요소

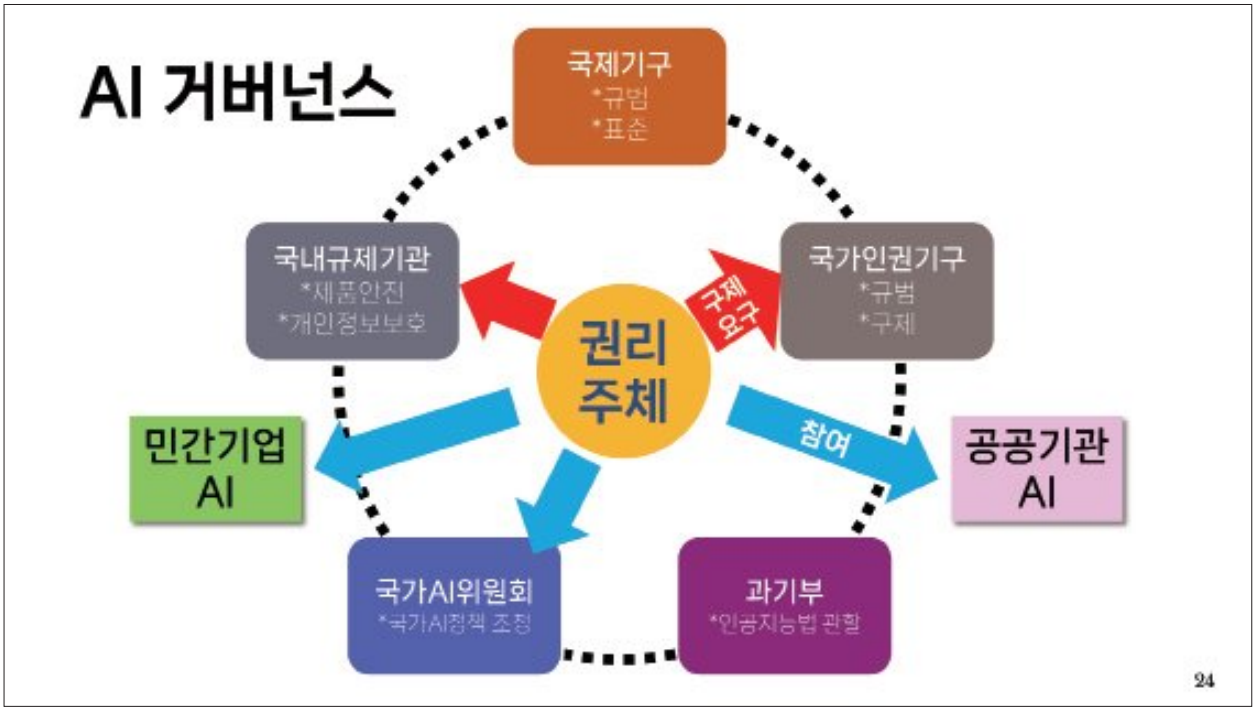
AI 인권기반 접근의 규범화 및 법제도 마련

- 국가와 기업의 인권법적 의무 이행
- AI의 투명성과 책무성 구현

영향받는 사람과 AI 거버넌스 조성

- 참여 거버넌스
- 구제 거버넌스 *인권기구

23


AI 거버넌스
국제기구
*규범
*표준
국내규제기관
*제품안전
*개인정보보호
국가인권기구
*규범
*구제
구제 요구
권리 주체
민간기업 AI
참여
공공기관 AI
국가AI위원회
*국가AI정책 조정
과기부
*인공지능법 관할
24


감사합니다
25

AI Human Rights Governance: Human Rights-Based Approach and Affected Persons

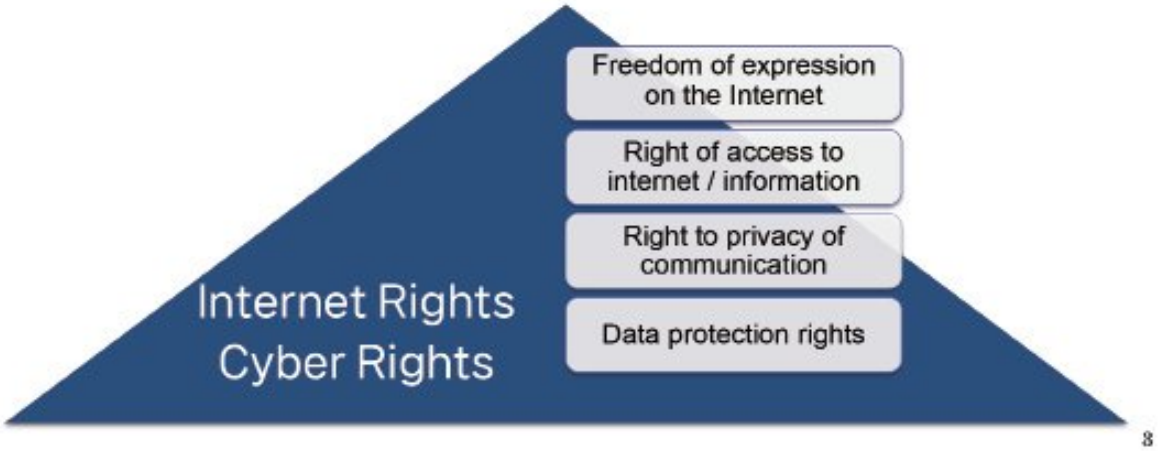
CHANG Yeo-kyung | Executive Director, Institute for Digital Rights

The Evolution of Digital Rights

2

The process of Discovering Rights-Holders and Identifying Rights

Demands that Began with the Rights of "Internet Users"



3

Digital Transformation

- Human activities, objects, and society as a whole—previously less related to digital technology are now interconnected through networks. From labor to leisure, all aspects of life are conducted on the basis of digital networks.

- The demand has expanded from the rights of internet users to the guarantee of universal digital rights for all indivisuals.

4

Identification of Information Rights by National Human Rights Institution

National Human Rights Commission of Korea
(Information Rights Report, 2013)

"Information rights are the fundamental rights that ensure human dignity is not harmed by the value of digitized information collected, processed, distributed, and utilized, and that it can be used freely and without discrimination."

5

Demand for Explicit Recognition of Information Rights

Constitutional Amendment Proposal by the President (2018)

- All people have the rights to know
- All people have the rights to have their personal information protected and to control its processing
- The state shall strive to prevent and correct the harms caused by the monopoly and disparity of information.
- Freedom of the press, publication, and other forms of expression shall be guaranteed, and permits or censorship for these shall be prohibited.

6

The Age of AI and its impact on Human Rights

Report of the UN Special Rapporteur on Extreme Poverty and Human Rights (2019)

- First, there are many issues raised by determining an individual's rights on the basis of predictions derived from the behavior of a general population group.
- Second, the functioning of the technologies and how they arrive at a certain score or classification is often secret, thus making it difficult to hold governments and private actors to account for potential rights violations.
- Third, risk-scoring and need categorization can reinforce or exacerbate existing inequalities and discrimination.

7

Human Rights-Based Approach to AI

8

Ethics-Based Approach

Driven by
Big Tech and
Government

기업

One Year after "Asiloma AI Principle", Kakao Announced its 'Algorithm Ethics Guidelines'

입력 : 2018-02-02 13:20:49 수정 : 2018-02-02 13:33:51

과학기술정보통신부 www.msit.go.kr

보도자료

한국판뉴딜 디지털뉴딜

보도일시	2020. 12. 23.(수) 11:00 이후 부터 보도해 주시기 바랍니다.		
배포일시	2020. 12. 22.(화) 14:00	담당부서	인공지능기반정책과
담당과장	김경민(044-202-6270)	담 당 자	유주연 사무관(044-202-6276)

The Ministry of Science and ICT formulated the [AI Ethics Standards], emphasizing a people-centered approach.

9

"Voluntary" Compliance

Ministry of Science and ICT (AI Ethics Standards, 2020)

(Objectives and Guidelines)
① All members of society, ② in every field, ③ shall voluntarily comply with, and ④ pursue continuously evolving ethical standards.

- As a moral regulation rather than a binding "law" or "guideline," it is a self-regulatory framework that respects corporate autonomy, encourages the advancement of artificial intelligence technologies, and fosters ethical discourse that can respond flexibly to technological and social changes.

10

International Human Rights Bodies' Critique of **'Voluntary Ethics'**

Report of the UN Special Rapporteur on Freedom of Opinion and Expression (2018)

- The private sector's focus on and the public sector's push for ethics often imply resistance to human rights-based regulation.

Report of the UN Special Rapporteur on Extreme Poverty and Human Rights (2019)

- Big Tech operates in an almost human rights free-zone.

11

Demand for Remedies Regarding AI Decision-Making

Houston Independent School District - EVAAS Case

U.S. District Court, Southern District of Texas (2017)

When public employment-related decisions are made using a teacher evaluation algorithm procured from a private vendor, the court held that

"Due process must be observed."

➔ Termination of the use of **proprietary secret algorithms** in important public decision-making

12

Proposal for a Human Rights-Based Approach

Report of the UN Special Rapporteur on Freedom of Opinion and Expression (2018)

- Humans are progressively excluded from defining the objectives and outputs of an AI system, ensuring **transparency, accountability** and access to **effective remedy** becomes more challenging...
- **Companies and Governments to take steps to permit systems to be scrutinized and challenged from conception to implementation.** Human rights impact assessments are one tool that can demonstrate a commitment to addressing the human rights implications of AI systems

13

Proposal for a Human Rights-Based Response

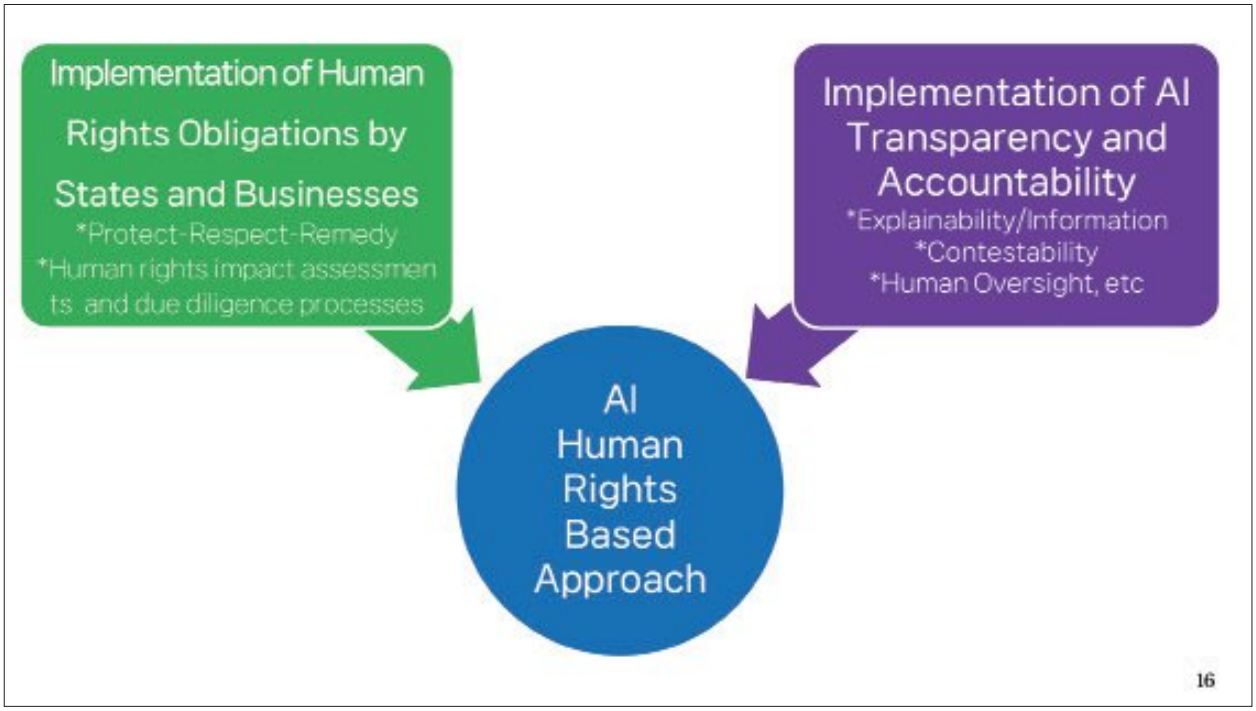
Report of the UN Secretary-General (2020)

- views people as **individual holders of rights**, empowers them
- and promotes a **legal and institutional** environment to enforce their rights
- seek **redress** for any human rights violations and abuses

14

Evolution of AI Human Rights Norms

NHRCK (2022) <Human Rights Guidelines for AI Development and Utilization>	COUNCIL OF EUROPE (2024) <Framework convention on artificial intelligence and human rights, democracy and the rule of law>
1. Respect for human dignity 2. Obligation of transparency and explanation 3. Guarantee of the right to self-determination 4. Prohibition of discrimination 5. Conduct of human rights impact assessments for AI 6. Classification of risk levels and establishment of related legislation	1. Human dignity and individual autonomy 2. Transparency and oversight 3. Accountability and responsibility 4. Equality and non-discrimination 5. Privacy and personal data protection 6. Reliability 7. Safe innovation


Implementation of Human Rights Obligations by States and Businesses
*Protect-Respect-Remedy
*Human rights impact assessmen ts and due diligence processes
Implementation of AI Transparency and Accountability
*Explainability/Information
*Contestability
*Human Oversight, etc
AI Human Rights Based Approach
16


Human Rights Governance in the Age of AI
17

Impacts of Automated Decision-Making

보도자료

보도일시	2020. 4. 14.(화) 조간(온라인 4. 13. 12:00)부터 보도해 주시기 바랍니다.		
배포일시	2020. 4. 13.(월) 09:00	담당부서	디지털사회기획과
담당과장	박민하(044-202-6130)	담 당 자	박 솔 사무관 (044-202-6135)

AI and Machine Learning to Drive Public Service Innovation: Ministry of Science and ICT Invests 20.7 Billion KRW

법제처 보도자료

배포 시점부터 보도해 주시기 바랍니다.

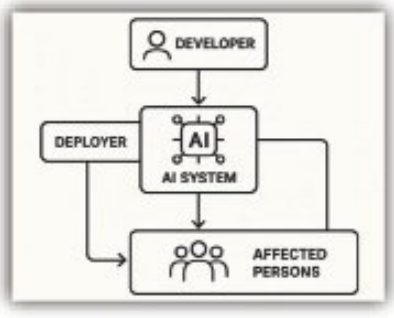
Article 20 (Automatic dispositions)
An administrative authority may impose a disposition using a fully-automated system (including systems in which **artificial intelligence technologies** are employed); provided, the same shall not apply to dispositions imposed at its discretion.

General Act on Public Administration Enacted and Enforced on March 23

18

Identifying Rights-Holders in the Age of AI

- 'Affected persons' refers to those whose life, physical safety, or fundamental rights are significantly impacted by artificial intelligence products or services (Article 2, Korea's Basic Act on AI)



19

Participation of the Affected Persons

UN Secretary-General (2020)

"States should create opportunities for **rights holders,** particularly those most affected or likely to suffer adverse consequences, to **effectively participate and contribute** to the development process, and facilitate targeted adoption of new technologies. **Through participation and inclusive consultation, States** can determine what technologies would be most appropriate and effective as they pursue balanced and integrated sustainable development with economic efficiency, environmental sustainability, inclusion and equity."

20

EU Example: Disclosure and Consultation with Affected Persons

Disclosure to Workers of High-Risk AI in the Workplace

Before putting into service or using a high-risk AI system at the workplace, deployers who are employers shall inform workers' representatives and the affected workers that they will be subject to the use of the high-risk AI system.

Fundamental Rights Impact Assessment as a Procedure for Participation and Consultation of Affected Persons

to collect relevant information necessary to perform the impact assessment, deployers of high-risk AI system, in particular when AI systems are used in the public sector, could involve relevant stakeholders, including the representatives of groups of persons likely to be affected by the AI system, independent experts, and civil society organisations in conducting such impact assessments and designing measures to be taken in the case of materialisation of the risks.

21

Remedies for Affected Persons

EU AI Act~ Council of Europe Framework Convention on AI

- **Provide information** on AI systems that have the potential to significantly affect human rights **to the competent authorities**, and, in certain cases, also to the affected persons.
- **Persons affected by AI-based decisions** shall have the possibility **to contest such decisions**, and, in certain cases, **to access information** about the system itself.
- Persons concerned shall have the right to **file a complaint** with the competent authorities.

22

Key Elements of AI Human Rights Governance

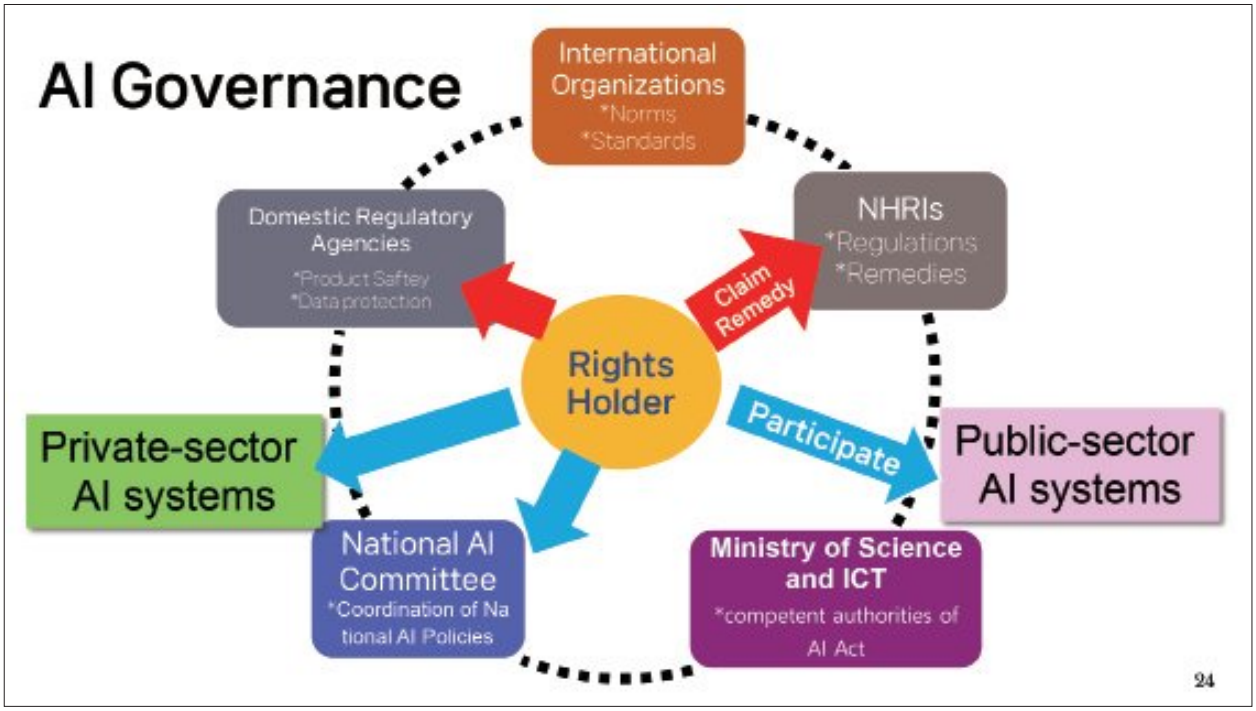
Norm-setting and legal frameworks for a human rights–based approach to AI

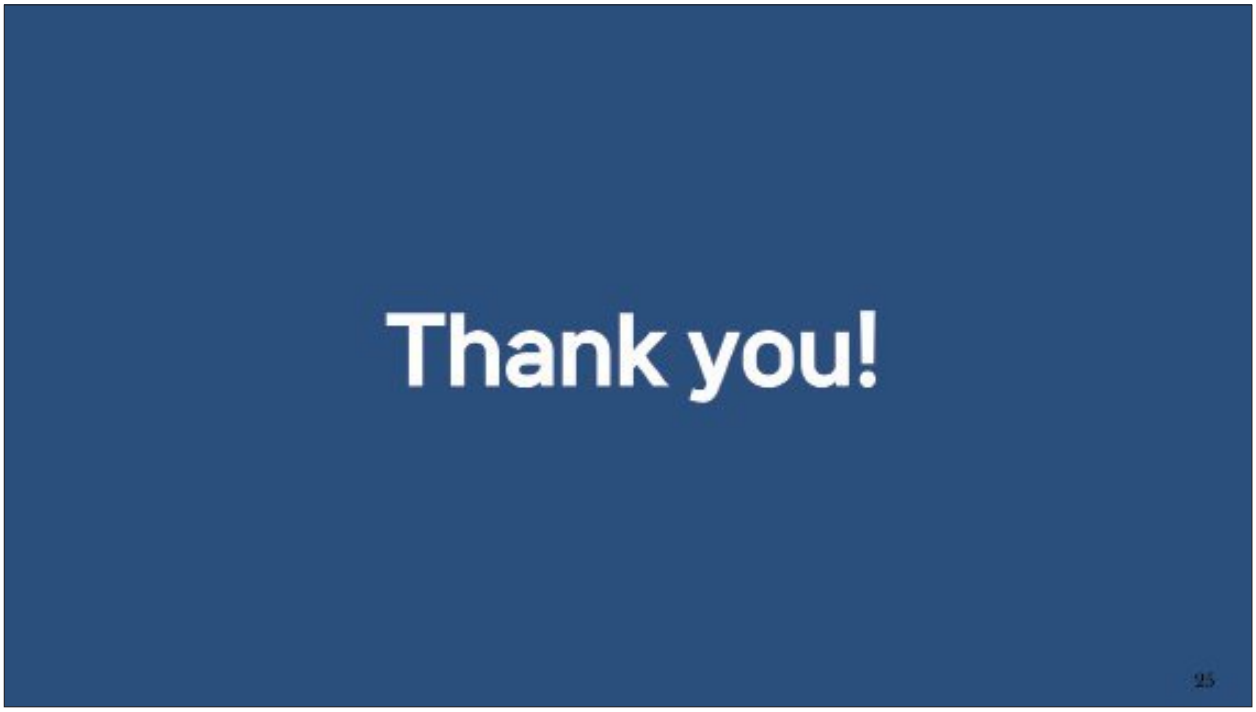
- Implementation of human rights obligations by states and businesses
- Ensuring transparency and accountability of AI systems

Affected Person and Building AI Governance

- Participatory Governance
- Remedy-focused Governance

23


AI Governance
International Organizations
*Norms
*Standards
Domestic Regulatory Agencies
*Product Saftey
*Data protection
NHRIs
*Regulations
*Remedies
Claim Remedy
Rights Holder
Participate
Private-sector AI systems
Public-sector AI systems
National AI Committee
*Coordination of National AI Policies
Ministry of Science and ICT
*competent authorities of AI Act
24


Thank you!
25

[발표 2 | Speaker 2]

AI 인권영향평가와 적용

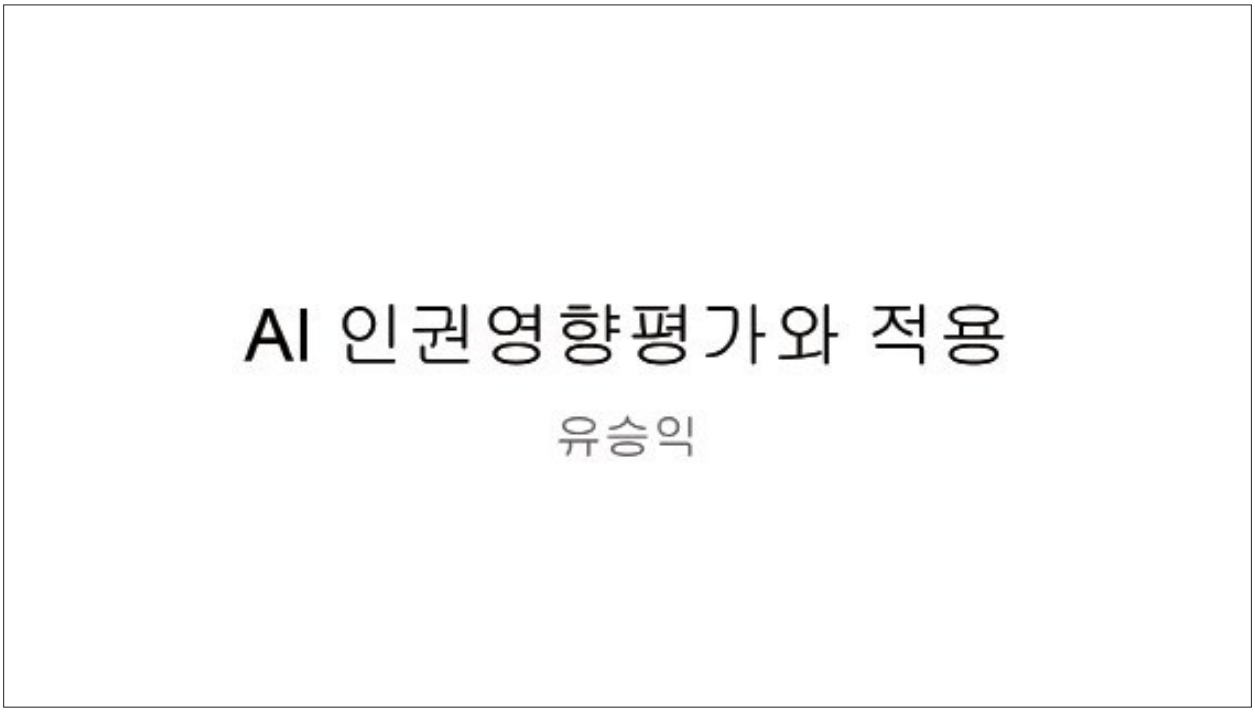
AI Human Rights Impact Assessment and Application

유승익
YOO Seung-ik

명지대학교 법학과 객원교수
Guest Professor, Department of Law, Myongji University

AI 인권영향평가와 적용

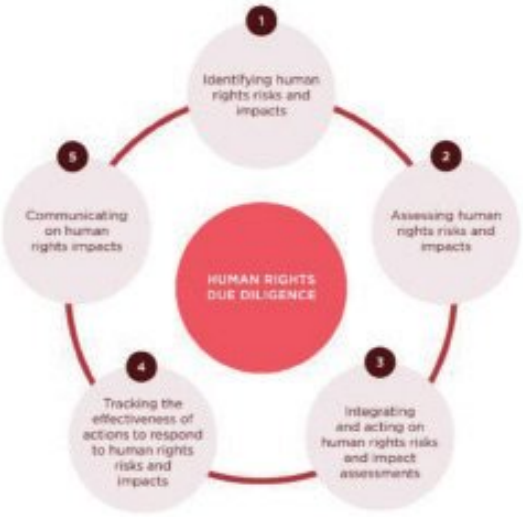
유승익 | 명지대학교 법학과 객원교수



인공지능 인권영향평가란 무엇인가?

2011년 기업과 인권에 관한 이행지침(UNGPs)

- 보호, 존중, 구제 프레임워크
  - 국가의 인권보호의무
  - 기업의 인권존중책임
  - 효과적인 구제 수단
- 부정적 인권 영향을 식별, 방지, 완화하고 그에 어떻게 대처하는지를 설명하기 위해 기업은 인권실사 (HRDD) 실시해야 함
- **인권실사의 핵심 도구로서 인권영향평가**



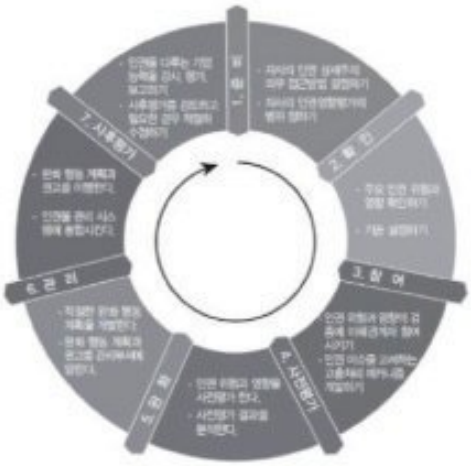
(인공지능에 의한) 많은 추론과 예측은, 사람들의 자율성과 자신의 정체성에 대한 세부사항을 확립할 권리를 포함하여, 프라이버시권의 향유에 깊은 영향을 미친다. 이는 또한 사상과 의견의 자유에 대한 권리, 표현의 자유, 공정한 재판 관련 권리 등 다른 권리에도 많은 문제를 야기한다.

인공지능 시스템의 설계, 개발, 배치, 판매, 구입, 운영의 **수명주기 전반**에 걸쳐 체계적으로 인권실사를 수행한다. 그 **인권 실사의 핵심 요소**는 **정례적이고 포괄적인 인권영향평가**여야 한다.

– 유엔 인권최고대표 <디지털 시대 프라이버시권(2021)>

인권영향평가(Human Rights Impact Assessment, HRIA)

- 사업과정, 정책, 입법, 프로젝트 등이 **인권에 미치는 영향을 측정하고 평가**하는 도구
- 국가, 기업 등이 시행 추진하는 사업과정이나 정책 등에서 인권에 미치는 부정적 영향을 식별, 방지, 완화하고 긍정적인 영향을 장려하기 위해 정책이나 사업 등의 계획과 활동이 인권의 실현과 보호에 부합하는지 평가하고 검토하는 것



국가인권위원회의 인권영향평가 권고

"39. 국가는 인공지능의 개발과 활용에 있어서 인권침해와 차별의 가능성 및 정도, 영향을 받는 당사자의 수, 사용된 데이터의 양 등을 고려하여 **공공기관 및 민간기업을 대상으로 인권영향평가를 실시**하여야 합니다."

– 국가인권위원회 <인공지능 개발과 활용에 관한 인권 가이드라인>(2022.4.)

"5. **인권영향평가 제도를 도입**하여 인공지능 개발·출시 전 인권영향평가를 실시하도록 하고, 출시 후 기능 수정 및 활용 범위 변경 시 재평가를 하도록 할 것"

– 국가인권위원회 「인공지능산업 육성 및 신뢰 기반 조성 등에 관한 법률안」에 대한 의견표명

국내 인공지능 평가도구

- 개인정보보호위원회, 인공지능(AI) 개인정보보호 자율점검표(2021.5.)
- 과학기술정보통신부, 신뢰할 수 있는 인공지능 실현전략(2021.5.)
- 금융위원회, <5大 금융분야 AI 개발·활용 안내서>(2022.8)
- 과학기술정보통신부, 한국정보통신기술협회(TTA), <신뢰할 수 있는 인공지능 개발 안내서 4종(2024. 3.)

인공지능 인권영향평가, 왜 필요한가?

- 인공지능 시스템의 불투명성과 파급효과
  - 사후적 피해구제의 어려움
- 인공지능 시스템의 중대한 인권 영향
  - 사상의 자유, 표현의 자유, 공정한 재판받을 권리 등에 심각한 영향
- 기존 각종 영향평가제도로 대응 역부족
  - 평가기준, 평가대상, 평가시기 및 주기, 평가주체 등 일관된 제도 필요
- 국제적 추세
  - 윤리기준을 넘어선 인공지능 관련 영향평가 도구 제도화 추세

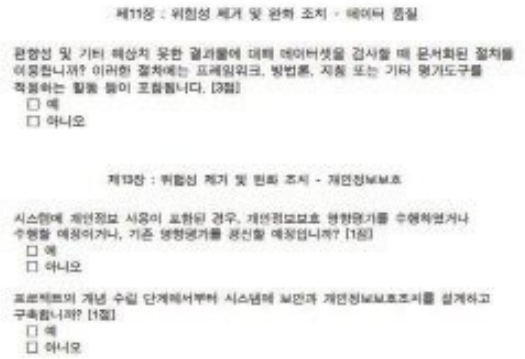
해외의 인공지능 영향평가

유럽연합 집행위원회, 신뢰할 수 있는 인공지능 평가 목록(2020.7.)

- 인공지능 영향평가의 초기 형태
- 평가목록 → 7대요구사항, 140여개 문항
  - 인간행위자와 감독
  - 기술적 견고성과 안정성
  - 프라이버시 및 데이터 거버넌스
  - 투명성
  - 다양성, 차별금지, 공정성
  - 사회, 환경적 복지
  - 책무성
- 기본권 영향평가 수행 제안
- 인공지능법(안)(2021.4.)으로 발전

캐나다 알고리즘영향평가

- 자동화된 의사결정 훈령(2019)에서 알고리즘 영향평가 도입
- 위험기반 접근법 : 4단계 위험수준 → 위험수준이 높을수록 높은 요구사항 부과
- 의사결정 알고리즘 사용하는 공공기관 의무적 실시
- 해당 기관이 온라인에서 수행
- 이해관계자 참여, 평가 결과 공


제11장 : 위험성 제거 및 완화 조치 - 데이터 품질

편향성 및 기타 예상치 못한 결과물에 대해 데이터셋을 검사할 때 문서화된 절차를 이용합니까? 이러한 절차에는 프레임워크, 방법론, 지침 또는 기타 평가도구를 적용하는 활동 등이 포함됩니다. [3점]
☐ 예
☐ 아니오

제13장 : 위험성 제거 및 완화 조치 - 개인정보보호

시스템에 개인정보 사용이 포함된 경우, 개인정보보호 영향평가를 수행하였거나 수행할 예정이거나, 기존 영향평가를 갱신할 예정입니까? [1점]
☐ 예
☐ 아니오

프로젝트의 개념 수립 단계에서부터 시스템에 보안과 개인정보보호조치를 설계하고 구축합니까? [1점]
☐ 예
☐ 아니오

영국 NMIP 알고리즘영향평가

- 2021년, 에이다 러브레이스 연구소가 NHS AI Lab 지원받아 개발
- 보건의료 분야 국가의료이미지플랫폼(National Medical Imaging Platform, NMIP)에 특화된 알고리즘 영향평가

2a 이 프로젝트가 특정 커뮤니티에 대한 불평등 또는 불법적인 차별의 생성 또는 악화로 이어질 수 있습니까? 예를 들어, 치료에 대한 차별적 접근을 악화시키면서? 편향 및 공정성을 평가하거나 모니터링하기 위한 현재 계획에서 간과할 수 있는 것은 무엇입니까?

성찰적 수행	참여 워크숍	종합

2b 귀하의 프로젝트는 동의와 자율성을 어떻게 고려합니까? 감시 증가와 관련된 위험이 있습니까? 예를 들어, 시스템의 의도된 수혜자에게 시스템 사용에 대해 어떻게 알립니까? 이 시스템은 감시가 증가하는 것으로 해석될 수 있습니까?

성찰적 수행	참여 워크숍	종합

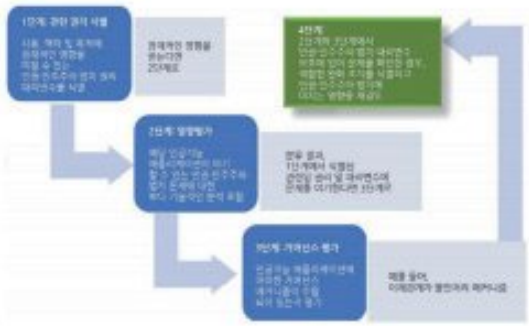
미국의 AI 영향평가(OMB)

- AI 사용에 대한 거버넌스, 혁신 및 리스 관리 강화 공문(M-24-10) : 정부기관들이 안전과 권리에 영향을 미치는 AI로부터 위험을 관리하기 위한 최소한의 조치를 취할 것을 권고
- 정부기관들은 안전에 영향을 미치는 AI, 권리에 영향을 미치는 AI에 대해 AI영향평가를 포함한 최소한의 위험관리 조치를 적용해야 함(기한: 2024.12.1. 1년 연장 가능(소명 필요))
- 정부기관은 AI영향평가 주기적 업데이트, AI 수명주기 전반에 걸쳐 활용해야 하고, 최소 사항은 문서화해야 함
- 트럼프 2기 행정부, America's AI Action Plan(2025.7) 발표

유럽평의회, 인권 민주주의 법치 영향평가

Human Rights, Democracy, and Rule of Law Impact Assessment of AI, HRDRIA(2021.5.)

- 지정학적, 사회적, 경제적 맥락 중시
- 인공지능 시스템 기반 기술 특성 고려: 범위, 신뢰성, 추적 가능성, 설명가능성, 처리 데이터 등
- 이해관계자 참여



덴마크 디지털활동 인권영향평가

- AI 등 디지털활동, 제품 및 서비스의 위험 특성을 평가하고 해결하기 위한 지침
- 인권기반 접근법 적용
- 인권영향평가의 다섯 단계
- 심각도 평가의 지표 제시 → 범위, 규모, 회복불가능성

덴마크 디지털활동 인권영향평가 : 심각도 평가 지표

항목	지표	심각도	비고
범위	영향 영역 전체 인구의 20% 이상 또는 식별된 집단의 50% 이상	A	인권 관점은 특정한 개인들이 함유하고 행사하는 인권과 자유를 강조한다. 따라서 범위(영향을 받는 사람의 수)를 고려할 때 절대적인 숫자만이 아니라 영향을 받는 개별 이용자 및 기타 권리주체가 누구인지 보다 정확하게 검토한다. 일부 영향은 수치적으로는 작을 수 있지만 비례적으로 더 큰 타격을 받는 특정 권리주체 집단에 편향될 수 있다. 예를 들어, 디지털 커뮤니케이션 플랫폼의 이용자 중 0.1%만이 영향을 받을 수 있지만 이것이 소수 종교인의 25%라면 후자가 전자보다 관련성이 더 높다. 집단 식별은 잠재적으로 영향을 받는 사람들(아성 이용자 또는 남성 이용자 등)을 세분화하는 상황별 방법론이 될 수 있다.
	영향 영역 전체 인구의 10% 이상 또는 식별된 집단의 10-50%	B	
	영향 영역 전체 인구의 5% 이상 또는 식별된 집단의 10% 미만	C	

- 평가지표
  - '범위(scope)'
  - '규모(scale)'
  - '회복불가능성(irremediability)'

항목	심각도
범위	A: 판결만을 보조하는 고도로 정확한 알고리즘에 경우 영향을 받는 사람의 절대적인 수는 적을 수 있지만, 사례를 자세히 살펴본 결과 주로 영향을 받는 사람들이 선주민으로 나타났고 취약 집단으로 식별되었다.
규모	A: 공정한 재판을 받을 권리에 대한 영향으로 상당한 규모의 인권영향이 있다. 알고리즘에 지원으로 이루어진 판결을 받은 모든 사람에게 적용된다. 그러나 영향을 받는 선주민과 관련한 영향은 훨씬 더 크며, 이는 그 집합이 본질적으로 취약하기 때문이다.
회복불가능성	B: 판결의 실격에 따라 심각을 판정이 구제되지 못할 수도 있다. 예를 들어, 이러한 작업 기하가 세션식이나 평생 영향을 미칠 수 있다. 설은 오랫동안 참된 관계에 영향을 끼칠 수 있으며, 수년간 가족생활을 영위할 수 없으며, 반드시 회복될 수 있는 것이 아니다. 그러나 영향의 일부는 시설적 관점을 변경하여 구제될 수 있다.
종합 평가	높은 심각도의 영향으로 간주될 수 있다. 이는 특히 취약 집단에 영향을 미치는 지속적인 영향이며 일부 사례(더도한 정책)는 예를 들어 건강과 생계 및 가족 생활에 대한 권리와 관련된 영향으로 인해 구제될 수 없다.

덴마크 디지털활동 인권영향평가 : 10개 핵심요소

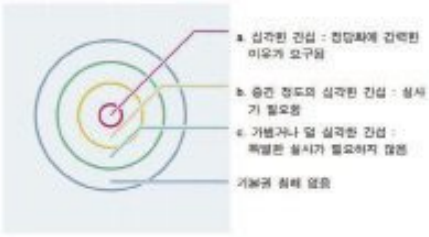
인권영향평가 절차 및 내용의 10개 핵심요소

구분	핵심 요소	설명
절차	① 참여	영향을 받았거나 잠재적으로 영향을 받을 권리주체의 유의미한 참여가 영향평가 절차의 모든 단계(데이터 수집 및 맥락 분석, 영향 분석, 영향 예방, 완화 및 개선, 보고 및 평가 등)에 반영된다.
	② 차별 금지	참여 및 협의 절차에서 포용적이고 심인지적이며 취약하고 소외될 위험이 있는 개인 및 집단의 요구를 고려한다.
	③ 역량 강화	취약하고 소외될 위험이 있는 개인 및 집단의 유의미한 참여를 보장하기 위해 이들에 대한 역량 강화를 수행한다.
	④ 투명성	영향평가 절차는 영향을 받았거나 잠재적으로 영향을 받을 수 있는 권리주체를 적절히 참여시키기 위해 가능한한 투명해야 하며, 권리주체나 기타 참여자(인권단체 및 언론 활동가 등)의 안전과 안녕에 위험을 초래하지 않는다. 영향평가 결과를 적절히 공개한다.
	⑤ 책무성	영향평가팀은 인권 전문가의 지원을 받으며 영향평가, 예방, 완화, 관리에 대한 역할과 책임을 부여하고 적절한 자원을 제공한다. 영향평가는 권리주체의 자격과 관련 의무주체의 의무와 책임(예: 개발자, 디지털 제품 및 서비스를 구입하는 기업, 이를 사용하거나 작동하는 기업 또는 정부 기관)을 파악한다.

구분	핵심 요소	설명
내용	⑥ 기준	인권 기준으로 영향평가의 기준을 구성한다. 영향 분석, 영향 심각도 평가 및 완화 조치 설계는 국제 인권 기준 및 원칙에 따라 수행된다.
	⑦ 영향 범위	평가는 사업이 야기하거나 기여한 실제적 및 잠재적 영향을 식별한다. 평가는 또한 기업의 운영, 제품 및 서비스 또는 사업 관계(계약 또는 비계약)를 통해 사업에서 직접적으로 관련된 영향을 고려한다. 평가는 누적적 영향 및 기존 문제를 분석한다.
	⑧ 영향의 심각도 평가	영향은 인권에 미치는 결과의 심각도에 따라 다루어진다. 여기에는 특정 영향의 범위, 규모 및 회복불가능성에 대한 검토가 포함되며, 권리주체 및 그 정당한 대리인의 견해를 고려한다.
	⑨ 영향 완화 조치	모든 인권영향이 다루어져야 한다. 영향을 다루기 위한 조치의 우선 순위를 정해야 하는 경우, 인권영향의 심각도가 핵심 기준이다. 식별된 영향을 해결할 때는 '방지-감소-회복-보상' 의 완화 계층 구조를 따른다.
	⑩ 구체수단 접근	영향을 받는 권리주체가 디지털활동, 제품 또는 서비스는 물론, 영향평가 절차 및 그 결과에 대한 진정을 제기할 수 있는 방법이 있어야 한다. 기업이 영향평가 및 관리에서 영향을 받는 권리주체를 위한 구체수단 접근 권한을 보장하거나 협력한다.

네덜란드 기본권 알고리즘영향평가(2022)

- AI 시스템을 개발하거나 도입하는 초기 단계에서 논의하고 해결해야 할 인권 쟁점에 대한 질의로 구성



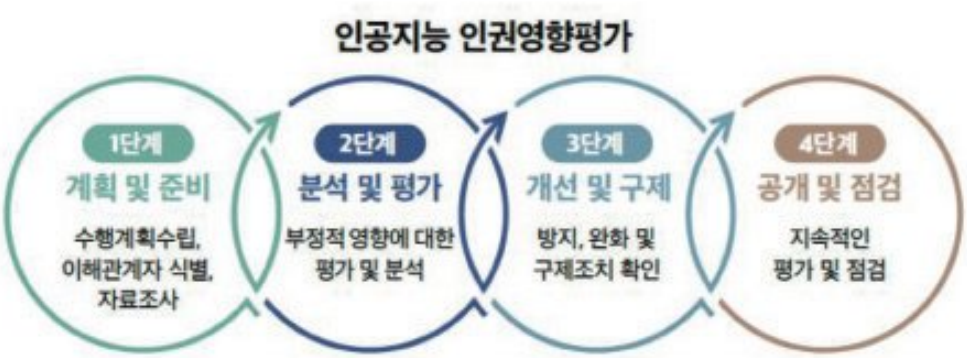
인공지능 인권영향평가 도구

국가인권위원회의 인공지능 인권영향평가 도구

인공지능 인권영향평가 도구 개요

- 평가대상
  - 공공기관이 도입(개발 및 조달)하는 모든 인공지능 + '고위험 인공지능'('금지대상 인공지능'은 평가대상에서 제외)
- 평가시기
  - 시스템 개발 전 계획 단계 또는 시행 전 단계 평가가 원칙
  - (평가 결과 일정 기준 이상의 구체적 위험이 확인된 경우) 정기적, 사후적 평가를 통한 지속적 관리
- 평가 수행 주체
  - 개발 및 활용 기관의 내부 또는 제3의 기관

인공지능 인권영향평가의 절차



평가 문항 구성

- 4단계 구성(각 단계별 질의 및 질의 취지에 대한 설명)
- 대부분 객관식으로 구성(일부 주관식)
- 체크리스트의 경우, 관련 내용 설명 요구
- 문항 구성
  - 개인정보보호 및 데이터 관리, 알고리즘의 신뢰성, 차별금지, 설명가능성과 투명성 등에 대한 영향 분석 및 식별
  - 위험에 대한 방지, 완화 및 구제
  - 이해관계자의 참여
  - 평가 결과의 공개 및 평가에 대한 점검 등

(4) 차별금지

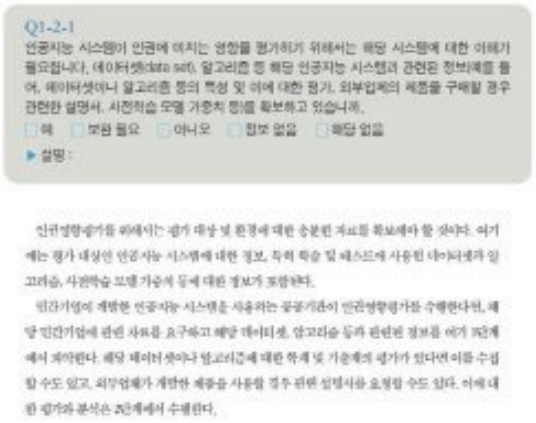
Q3-1-11
인공지능 시스템이 활용 과정에서 합리적인 이유 없이 인종, 종교, 장애, 나이, 학력, 직업, 출신 지역, 언어, 정치 성향, 신체조건, 외모, 피부색, 방탁, 성별, 성적 지향, 사회적 신분, 경제적 지위 등 개인과 집단의 특성에 따라 특정 집단에 대한 차별을 야기하거나 혹은 기존의 차별을 악화시킬 가능성이 있는지 검토하였습니까.

□예 □보완 필요 □아니오 □정보 없음 □해당 없음

▶ 설명 :

인공지능 인권영향평가의 절차

- **1단계: 계획 및 준비**
- 가. 인권영향평가 계획
  - 평가 대상 인공지능 시스템 및 이해관계자 파악을 위한 질의
- 나. 사전 조사
  - 인공지능 시스템 도입, 활용되는 분야나 사회의 맥락, 특성, 이해관계자와의 협의 등을 위한 질의


Q1-2-1
인공지능 시스템이 인권에 미치는 영향을 평가하기 위해서는 해당 시스템에 대한 이해가 필요합니다. (데이터셋(data set), 알고리즘 등 해당 인공지능 시스템과 관련된 정보제공을 통해, 데이터셋이나 알고리즘 등의 특성 및 이에 대한 평가, 외부업체의 제품을 구매할 경우 관련한 설명서, 사전학습 모델 가중치 등을 확보하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
▶ 설명 :

인권영향평가를 위해서는 평가 대상 및 환경에 대한 충분한 자료를 확보해야 할 것이다. 여기에는 평가 대상인 인공지능 시스템에 대한 정보, 특히 학습 및 테스트에 사용된 데이터셋과 알고리즘, 사전학습 모델 가중치 등에 대한 정보가 포함된다.
인간기업이 개발한 인공지능 시스템을 사용하는 공공기관이 인권영향평가를 수행한다면, 해당 인간기업에 관련 자료를 요구하고 해당 데이터셋, 알고리즘 등과 관련된 정보를 여기 1단계에서 파악한다. 해당 데이터 셋이나 알고리즘에 대한 학계 및 기술계의 평가가 있다면 이를 수집할 수도 있고, 외부업체가 개발한 제품을 사용할 경우 관련 설명서를 요청할 수도 있다. 이에 대한 평가와 분석은 2단계에서 수행한다.

인공지능 인권영향평가의 절차

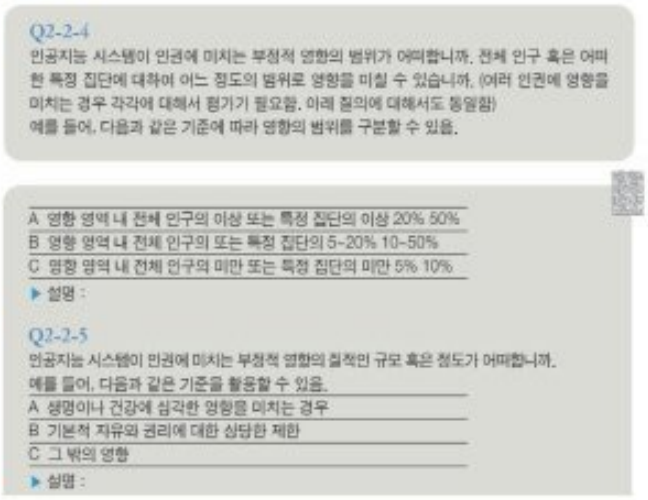
- **2단계: 분석 및 평가**
- 가. 인공지능 기술과 관련된 영향 분석 및 평가
  - 개인정보보호
  - 데이터 관리
  - 알고리즘의 성능과 신뢰성
  - 차별금지
  - 설명가능성과 투명성
  - 자동화 정도와 인간의 개입
  - 보안
  - 접근성
  - 라이선스


Q2-1-1
인공지능 시스템이 개인정보보호위원회 〈인공지능(AI) 개인정보보호 자율점검표〉의 모든 의무/권장 조항을 준수하고 있습니까.
☐ 예 ☐ 보완 필요 ☐ 아니오 ☐ 정보 없음 ☐ 해당 없음
▶ 설명 :

인공지능 시스템 개발 및 운영 과정에서 수집, 처리되는 개인정보는 개인정보 보호법에 따라 적법하게 처리되어야 한다. 개인정보 보호법의 준수 여부 확인을 위해 데이터 최소화의 원칙, 개인정보 처리의 법적 근거, 정보 주체의 권리 보장 여부, 개인정보에 대한 안전성 조치 등을 검토할 필요가 있다. 본 인권영향평가 도구에서는 개인정보보호위원회가 발표한 〈인공지능(AI) 개인정보보호 자율점검표〉의 요구사항을 준수하고 있는지 여부를 확인하도록 하였다.

인공지능 인권영향평가의 절차

- **2단계: 분석 및 평가**
- 나. 인권에 미치는 영향 및 심각도
  - (1) 영향을 받는 인권 : 예시 제공
  - (2) 인권에 미치는 영향의 심각도



Q2-2-4
인공지능 시스템이 인권에 미치는 부정적 영향의 범위가 어떠합니까. 전체 인구 혹은 어떠한 특정 집단에 대하여 어느 정도의 범위로 영향을 미칠 수 있습니까. (여러 인권에 영향을 미치는 경우 각각에 대해서 평가가 필요함. 아래 질의에 대해서도 동일함)
예를 들어, 다음과 같은 기준에 따라 영향의 범위를 구분할 수 있음.

A 영향 영역 내 전체 인구의 이상 또는 특정 집단의 이상 20% 50%
B 영향 영역 내 전체 인구의 또는 특정 집단의 5-20% 10-50%
C 영향 영역 내 전체 인구의 미만 또는 특정 집단의 미만 5% 10%
▶ 설명 :

Q2-2-5
인공지능 시스템이 인권에 미치는 부정적 영향의 질적인 규모 혹은 정도가 어떠합니까.
예를 들어, 다음과 같은 기준을 활용할 수 있음.
A 생명이나 건강에 심각한 영향을 미치는 경우
B 기본적 자유와 권리에 대한 상당한 제한
C 그 밖의 영향
▶ 설명 :

인공지능 인권영향평가의 절차

- **3단계: 개선 및 구제**
- 가. 방지
  - 인권침해 위험 방지 조치 식별
- 나. 완화
  - 인권 침해 위험 완화 조치 검토
- 다. 구제
  - 이의 제기절차 및 권리침해 구제 절차
- 라. 이해관계자와의 의견수렴 및 협의
  - 위험방지, 완화, 피해구제 방안에 대한 이해관계자와의 협의

- **4단계: 공개 및 점검**
- 가. 인공지능 시스템의 주요 요소의 공개
- 나. 인권영향평가 결과 공개
- 다. 사후 모니터링
- 라. 인권영향평가에 대한 점검
- 마. 인권영향평가의 재수행

AI 인권영향평가 관련 주요 과제

인공지능 기본법 제35조(고영향 인공지능 영향평가) ① 인공지능사업자가 고영향 인공지능을 이용한 제품 또는 서비스를 제공하는 경우 사전에 사람의 기본권에 미치는 영향을 평가(이하 "영향평가"라 한다)하기 위하여 노력하여야 한다.

- 평가 대상의 범위 : 고위험(고영향) 인공지능의 정의와 범위
- 평가 주체 : 인공지능사업자(개발사업자 + 이용사업자)
- 노력의무 조항 : 자율규제 vs. 의무규제 → 실효성 문제(우선고려)
- 평가 시기 : 사전 영향평가 → 사후적, 정기적 평가 필요
- 영향평가 과정 및 공개 과정에서 기업 영업비밀, 지적재산권 침해 가능성
- 다른 평가체계와의 관계 설정 문제

감사합니다

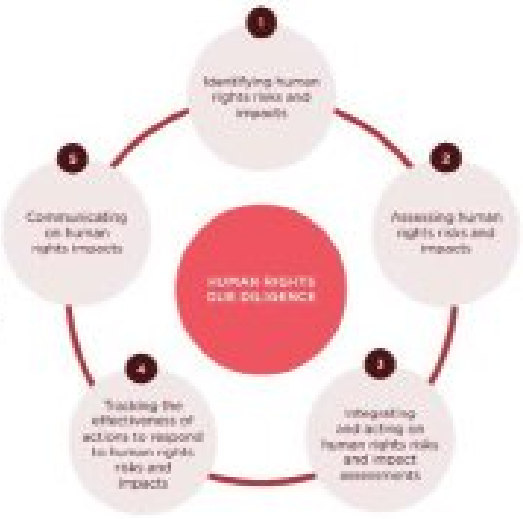
AI Human Rights Impact Assessment and Application

YOO Seung-ik | Guest Professor, Department of Law, Myongji University

What is an AI human rights impact assessment?

2011 UN Guiding Principles on Business and Human Rights (UNGPs)

- Protect, Respect and Remedy Framework
  - State duty to protect human rights
  - Corporate responsibility to respect human rights
  - Access to effective remedies
- Businesses should conduct Human Rights Due Diligence (HRDD) to identify, prevent, mitigate and account for how they address their adverse human rights impacts.
- **Human Rights Impact Assessment (HRIA) is a core tool of HRDD.**



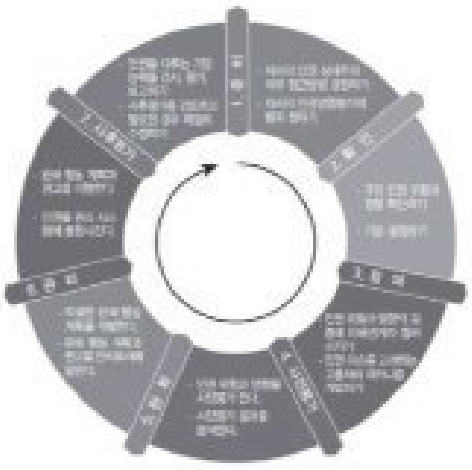
Many inferences and predictions deeply affect the enjoyment of the right to privacy, including people's autonomy and their right to establish details of their identity. They also raise many questions concerning other rights, such as the rights to freedom of thought and of opinion, the right to freedom of expression, and the right to a fair trial and related rights

Human rights due diligence must be carried out systematically throughout the entire life cycle of AI systems—from design, development, deployment, sale, purchase, to operation. A central element of such due diligence should be regular and comprehensive human rights impact assessments.

– UN High Commissioner for Human Rights & Privacy Rights in the Digital Age (2021)

Human Rights Impact Assessment (HRIA)

- A tool to measure and evaluate the impact of business processes, policies, legislation, and projects on human rights
- An evaluation of whether state or corporate policies and projects are consistent with the realization and protection of human rights, by identifying, preventing, and mitigating negative impacts while encouraging positive contributions..



National Human Rights Commission of Korea's Recommendations on the Human Rights Impact Assessment

"39. The State shall conduct **human rights impact assessments on public institutions and private enterprises** in the development and use of AI, taking into account factors such as the likelihood and extent of human rights violations and discrimination, the number of individuals affected, and the amount of data used.."

– National Human Rights Commission of Korea, *Human Rights Guidelines on the Development and Use of Artificial Intelligence*, (April 2022)

"5. **Establish a human rights impact assessment system** to ensure that assessments are conducted prior to the development and deployment of AI, and that reassessments are carried out when functions are modified or the scope of use is expanded after deployment."

– National Human Rights Commission of Korea, *Opinion on the Draft Framework Act on the Development of Artificial Intelligence and the Establishment of Foundation for Trustworthiness*

National artificial intelligence evaluation tools

- Personal Information Protection Commission, AI Personal Information Protection Self-Checklist (May 2021)
- Ministry of Science and ICT (MSIT), Strategy for Ensuring Trustworthy AI (May 2021)
- Financial Services Commission, Guidelines for AI Development and Utilization in the Five Major Financial Sectors (August 2022)
- Ministry of Science and ICT (MSIT) and Telecommunications Technology Association (TTA), Four Guidelines for Developing Trustworthy AI (March 2024)

Why Are AI Human Rights Impact Assessments Necessary?

- Opacity and Ripple Effects of Artificial Intelligence Systems
  - Difficulties in providing remedies after harm occurs
- Significant human rights implications of AI systems
  - Serious effects on freedom of thought, freedom of expression, and the right to a fair trial
- Insufficiency of existing impact assessment mechanisms
  - Lack of consistent standards for assessment scope, timing, frequency, and responsible actors
- International Trends
  - Trend toward institutionalising AI-related impact assessment tools beyond ethical standards

AI Impact Assessments Abroad

European Commission, *Assessment List for Trustworthy AI* (July 2020)

- Initial form of artificial intelligence impact assessment
- Assessment List → 7 Key Requirements, Over 140 Questions
  - human agency and oversight
  - technical robustness and safety
  - privacy and data governance
  - transparency
  - diversity, non-discrimination and fairness
  - environmental and societal well-being and
  - Accountability
- Proposal to conduct a fundamental rights impact assessment
- Evolved into the AI Act (draft, April 2021)

Canada's Algorithmic Impact Assessment (AIA)

- Introduced under the Directive on Automated Decision-Making (2019)
- Risk-based approach: four levels of risk → higher risk = stronger requirements
- Mandatory for public institutions using decision-making algorithms
- Conducted online by the implementing institution
- Stakeholder participation and publication of assessment results

Chapter 11: Risk Elimination and Mitigation Measures – Data Quality

Do you use documented procedures when examining datasets for bias and other unexpected outcomes? Such procedures may include applying frameworks, methodologies, guidelines, or other evaluation tools. (3 points)

☐ Yes
☐ No

Chapter 13: Risk Elimination and Mitigation Measures – Personal Data Protection

If the system involves the use of personal data, have you conducted, planned to conduct, or planned to update an existing personal data protection impact assessment? (1 point)

☐ Yes
☐ No

At the project's concept development stage, do you design and implement security and personal data protection measures in the system? (1 point)

☐ Yes
☐ No

UK NMIP Algorithm Impact Assessment

- Developed in 2021 by the **Ada Lovelace Institute** with support from the **NHS AI Lab**
- Tailored Algorithmic Impact Assessment for the **National Medical Imaging Platform (NMIP)** in the healthcare sector

2.a Could this project result in inequality or unlawful discrimination against a particular community, or could it worsen such impacts? For example, could it exacerbate discriminatory access to medical treatment? In your current plan, what measures can be identified to assess or monitor bias and fairness?

Scoping	Participation Workshop	Integration

2.b How does your project consider consent and autonomy? Are there risks related to increased surveillance? For example, how will the intended user of the system know about the use of the system? Could the system be interpreted as leading to an increase in surveillance?

Scoping	Participation Workshop	Integration

U.S. AI Impact Assessment (OMB)

- Memorandum on Strengthening AI Governance, Innovation, and Risk Management (M-24-10): Recommends that federal agencies take minimum measures to manage risks from AI systems that affect safety and rights.
- Federal agencies must apply minimum risk management measures, including AI impact assessments, for AI systems that affect safety and rights (deadline: December 1, 2024, with the possibility of a one-year extension upon justification).
- Federal agencies shall periodically update AI impact assessments, apply them throughout the AI lifecycle, and document minimum requirements.
- Under the Trump administration's second term, America's AI Action Plan was announced (July 2025).

Human Rights, Democracy and Rule of Law Impact Assessment of AI

Human Rights, Democracy, and Rule of Law Impact Assessment of AI, HRDRIA (May 2021)

- Emphasis on geopolitical, social, and economic contexts
- Consideration of the technical characteristics of AI systems: scope, reliability, reliability, traceability, explainability, data processing, etc.
- Stakeholder participation

Denmark – Human Rights Impact Assessment of Digital Activities

- Guidelines for assessing and addressing the risk characteristics of digital activities, products, and services, including AI
- Application of a human rights-based approach
- Five steps of human rights impact assessment
- Indicators for assessing severity→ scope, scale, irremediability

Denmark – Human Rights Impact Assessment of Digital Activities: Severity Assessment Indicators

Item	Indicator	Severity	Notes
Scope	Impact affects 20% or more of the total population in the area of influence, or 50% or more of a specific group	A	From a human rights perspective, the emphasis is placed on the rights and freedoms enjoyed and exercised by specific individuals. Accordingly, when considering scope (the number of people affected), it is necessary to examine not only the absolute number but also more precisely who the affected individuals and other rights holders are. Some impacts may appear economically small in many disproportionately affect particular groups of rights holders. For example, only 0.1% of users of a digital recommendation platform may be affected, but if this represents 20% of a minority religious group, the latter is of greater significance. Group identification can serve as a contextual methodology for disaggregating potentially affected populations (e.g., women, users of a state assist).
	Impact affects 10% or more of the total population in the area of influence, or 10–50% of a specific group	B	
	Impact affects 5% or more of the total population in the area of influence, or less than 10% of a specific group	C	

- Assessment Indicators
  - Scope
  - Scale
  - Irremediability

Item	Severity
Scope	D: In cases where algorithms are highly accurate and precise, the absolute number of people affected may be small, but detailed case analysis has shown that these persons impacted tend to be marginalized individuals or vulnerable groups.
Scale	A: There is a considerable recognized impact on the right to a fair trial. With the support of algorithms, all individuals who receive judgments are affected. However, the impact is much greater on defendants and suspects, as this impact is inherently discriminatory.
Irremediability	B: Depending on the nature of the judgment, the violations may not be fully remedied. For example, a wrong judgment may affect a person's lifelong employment opportunities, health, and social standing. Even if the judgment is overturned years later, the negative effects cannot be fully undone, meaning they are not completely remediable. Nevertheless, such impacts can vary depending on the details of the case.
Overall Assessment	A high level of severity of impact may be recognized. This is particularly true when there is a continuing impact on vulnerable groups. In some cases, where judgments are highly accurate, rights and related interests such as health, livelihood, and safety may still be restricted, resulting in irremediable harm.

Denmark's Digital Activities Human Rights Impact Assessment: 10 Key Elements

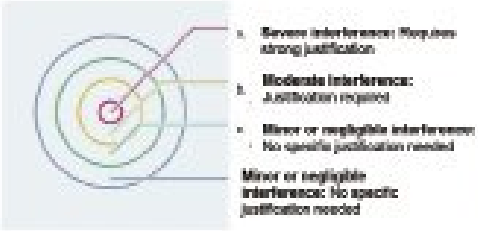
10 Core Elements of the Human Rights Impact Assessment Procedure and Content

Item	Core Elements	Explanation
Procedure	① Participation	Meaningful participation of rights holders who are affected or potentially affected shall be reflected in all stages of the impact assessment process (data collection and context analysis, impact analysis, prevention and mitigation, reporting and evaluation, etc.)
	② Non-Discrimination	Participation and consultation procedures should inclusively and substantively take into account the burdens of individuals and groups who are vulnerable or at risk of marginalization
	③ Empowerment	Measures should be taken to strengthen the capacities of vulnerable or marginalized rights-holders to ensure their meaningful participation.
	④ Transparency	The impact assessment process shall ensure that affected or potentially affected rights-holders are appropriately engaged, and its results are disclosed in a way that does not compromise their safety or human rights.
	⑤ Accountability	With the support of human rights experts, the impact assessment assigns roles and responsibilities for prevention, mitigation, remedy, and management, and ensures adequate resources. Accountability applies not only to rights-holders themselves, but also to duty-bearers (governments, companies), accountability mechanisms, providers of digital products and services, and the entities using them (including government agencies).

Item	Core Elements	Explanation
Content	⑥ Standards	Human rights standards constitute the basis for the assessment, impact analysis, severity assessment, and the design of mitigation measures shall be carried out in accordance with international human rights norms and principles
	⑦ Scope of Impact	The assessment identifies actual or potential impacts that the business causes or contributes to. It also considers impacts arising from the company's operations, products and services, or business relationships (direct or indirect). Cumulative impacts are also analyzed.
	⑧ Severity Assessment	Impacts are addressed according to their severity in terms of consequences for human rights. Severity factors include scope, scale, and irremediability, while also taking into account the perspectives of rights holders and their groups.
	⑨ Mitigation Measures	All human rights impacts must be addressed. When prioritization is necessary, severity of human rights impacts is the key criterion. Identified impacts shall be addressed through a mitigation framework of "prevent–reduce–restore–compensate."
	⑩ Access to Remedy	Rights holders must have access to remedy channels not only in relation to digital activities, products, or services, but also regarding the impact assessment procedure and its outcomes. Companies and governments must guarantee or cooperate to ensure effective access to remedy for rights holders.

Netherlands – Fundamental Rights and Algorithms Impact Assessment (FRAIA, 2022)

- A set of questions on human rights issues to be addressed at the early stage of AI system development or deployment.



AI Human Rights Impact Assessment Tool

National Human Rights Commission of Korea
AI Human Rights Impact Assessment Tool

Overview of AI Human Rights Impact Assessment Tool

- Scope of Assessment
  - All AI systems developed, procured, or adopted by public institutions, as well as "high-risk AI."("Prohibited AI" is excluded from the scope of assessment.)
- Timing or Assessment
  - In principle, assessments are to be conducted at the planning stage prior to system development or before implementation.
  - If the assessment identifies concrete risks above a certain threshold, continuous management shall be ensured through regular and ex-post assessments.
- Assessment Body
  - The assessment may be carried out either internally by the developing/implementing institution or by a third-party body.

Procedure of AI Human Rights Impact Assessment

Structure of Human Rights Impact Assessment Items

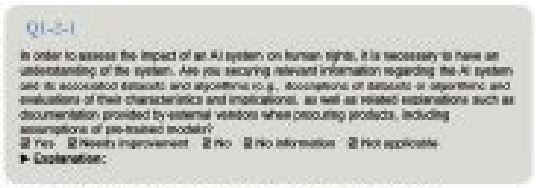
- 4-Stage structure (Each stage includes questions, along with an explanation of their rationale.)
- Mostly multiple-choice questions (with some open-ended items)
- Additional explanations are required for checklist.
- Set of Questions
  - Analysis and identification of impacts concerning personal data protection and data management, algorithm reliability, non-discrimination, explainability, and transparency
  - Prevention, mitigation, and remedy of risks
  - Stakeholder participation
  - Disclosure of assessment results and review of the evaluation process

(4) Non-Discrimination
Q2-1-11
In the process of using the AI system, has it been examined whether there is a possibility of causing discrimination against, or exacerbating existing discrimination of, certain individuals or groups based on characteristics such as race, religion, disability, age, educational background, occupation, place of birth, region, language, political orientation, physical condition, appearance, skin color, illness, gender, sexual orientation, social status, or economic standing?

☐ Yes ☐ Needs improvement ☐ No ☐ No information ☐ Not applicable

Procedures for AI Human Rights Impact Assessment

- **Step 1: Planning and Preparation**
- A. Planning the Human Rights Impact Assessment
  - Questions to identify AI systems and stakeholders subject to assessment
- b. Preliminary investigation
  - Questions to identify the AI system, areas of application, social context, characteristics, and consultation with stakeholders


Q1-2-1

In order to assess the impact of an AI system on human rights, it is necessary to have an understanding of the system. Are you securing relevant information regarding the AI system and its associated datasets and algorithms (e.g., assumptions on datasets or algorithms and evaluations of their characteristics and implications), as well as related explanations such as documentation provided by external vendors when procuring products, including assumptions of pre-trained models?

☑ Yes ☑ Needs improvement ☑ No ☑ No information ☑ Not applicable

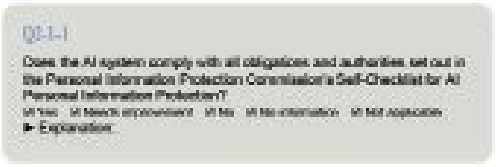
► Explanation:

To conduct a human rights impact assessment, sufficient information must be secured regarding the system subject to assessment and its environment. This includes information on the AI system to be assessed, particularly the datasets and algorithms used for training and testing, as well as assumptions of pre-trained models and their weights.

When a public institution conducts a human rights impact assessment of an AI system developed by a private company, it should request relevant materials from the company and obtain related information such as datasets and algorithms at this stage (Step 1). If the company has conducted its own internal or technical evaluation of the datasets or algorithms, such assessments may be collected. In addition, if products developed by external vendors are used, related documentation may be requested. Further analysis of this information is conducted in Step 2.

Procedures for AI Human Rights Impact Assessment

- **Stage 2: Analysis and Evaluation**
- A. Analysis and Evaluation of the Impacts Related to Artificial Intelligence Technology
  - Protection of personal data
  - Data Management
  - Algorithm Performance and Reliability
  - Non-discrimination
  - Explainability and Transparency
  - Degree of Automation and Human Involvement
  - Security
  - Accessibility
  - Licensing


Q1-1-1

Does the AI system comply with all obligations and authorities set out in the Personal Information Protection Commission's Self-Checklist for AI Personal Information Protection?

☑ Yes ☑ Needs improvement ☑ No ☑ No information ☑ Not applicable

► Explanation:

Personal data collected and processed in the course of developing and operating an AI system must be handled in accordance with the Personal Information Protection Act. To verify compliance with this Act, it is necessary to examine the principles of data minimization, legal grounds for personal data processing, guarantees of data subjects' rights, and the implementation of security measures for personal data. This human rights impact assessment tool is designed to ensure that compliance with the requirements of the *Self-Checklist for AI Personal Information Protection* published by the Personal Information Protection Commission is verified.

Procedures for AI Human Rights Impact Assessment

- **Stage 2: Analysis and Evaluation**
- b. Human Rights Impacts and Severity
  - (1) **Rights Affected:** with examples provided
  - (2) **Severity of Impacts on Human Rights**

Q2-2-4
What is the scope of the negative impact that the AI system may have on human rights? To what extent could it affect the entire population or certain specific groups? (If multiple human rights are affected, each must be assessed separately. The same applies to the following questions.)
For example, the scope of impact can be categorized according to the following criteria.

A. 20–50% or more of the total population in the affected area, or 20–50% or more of a specific group
B. 5–20% or 10–50% of the total population in the affected area, or of a specific group
C. Less than 5–10% of the total population in the affected area, or of a specific group
▶ **Explanation:**

Q2-2-5
What is the qualitative scale or degree of the negative impact that the AI system may have on human rights? For example, the following criteria may be applied:
A. Severe impact on life or health
B. Significant restriction on fundamental freedoms and rights
C. Other impacts
▶ **Explanation:**

Procedures for AI human rights impact assessment

- **Stage 3: Improvement and Remedy**

A. Prevention
Identify measures to prevent risks of human rights violations

b. Mitigation
Review of measures to mitigate risks of human rights violations

C. Remedy
Complaint Procedures and Remedies for Human Rights Violations

D. Consultation with Stakeholders
Consultation with stakeholders on measures for risk prevention, mitigation, and redress of harm

- **Stage 4: Disclosure and Review**

A. Disclosure of key elements of the artificial intelligence system

B. Disclosure of the results of the human rights impact assessment

C. Post-implementation monitoring

C. Review of the human rights impact assessment

E. Re-conducting the Human Rights Impact Assessment

Key tasks related to AI human rights impact assessment

Article 35 (Impact Assessment of High-Impact Artificial Intelligence) of the Framework Act on the Development of Artificial Intelligence and the Establishment of Foundation for Trustworthiness

① Where an AI provider offers products or services that utilize high-impact AI, the provider shall endeavor to conduct an assessment (hereinafter referred to as "Impact Assessment") in advance to evaluate the potential effects on fundamental rights.

- Scope of assessment: Definition and scope of high-risk (high-impact) AI
- Assessment Subjects: AI providers (both developers and users)
- Duty of Effort Provisions: Self-regulation vs. Mandatory Regulation → Effectiveness Issues (Priority Consideration)
- Timing of the assessment: Pre-assessment → Need for post-assessment and periodic assessments
- Possibility of infringement of trade secrets and intellectual property rights during the assessment process and disclosure process
- Issues related to the relationship with other evaluation systems

Thank you

[발표 3 | Speaker 3]

신기술과 인권에 대한 국가인권기구의 역할

The Role of NHRIs for the New Technology and Human Rights

넬레 루켄스

Nele ROEKENS

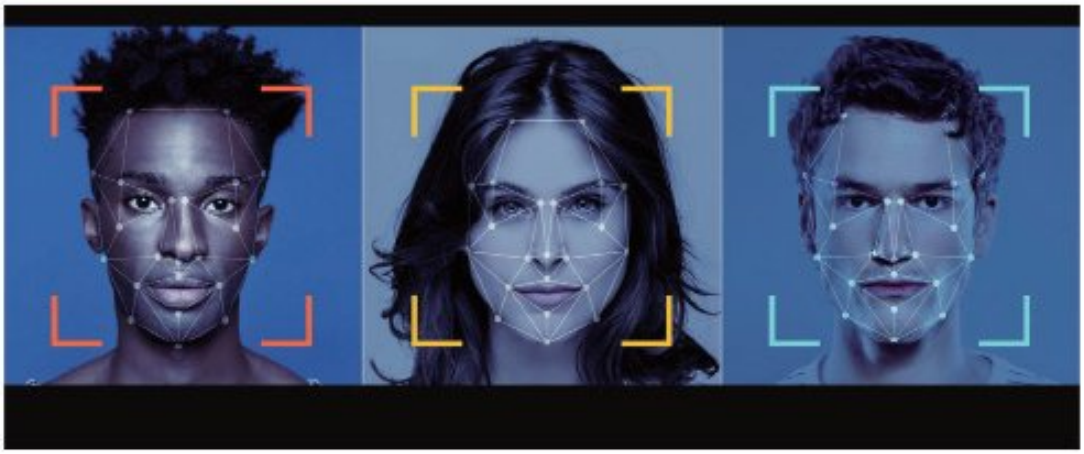
유럽 국가인권기구 연합 AI와 인권 실무그룹 의장

Chair of the Working Group on AI, ENNHRI

신기술과 인권에 대한 국가인권기구의 역할

넬레 루켄스 | 유럽 국가인권기구 연합 AI와 인권 실무그룹 의장

인권 과제

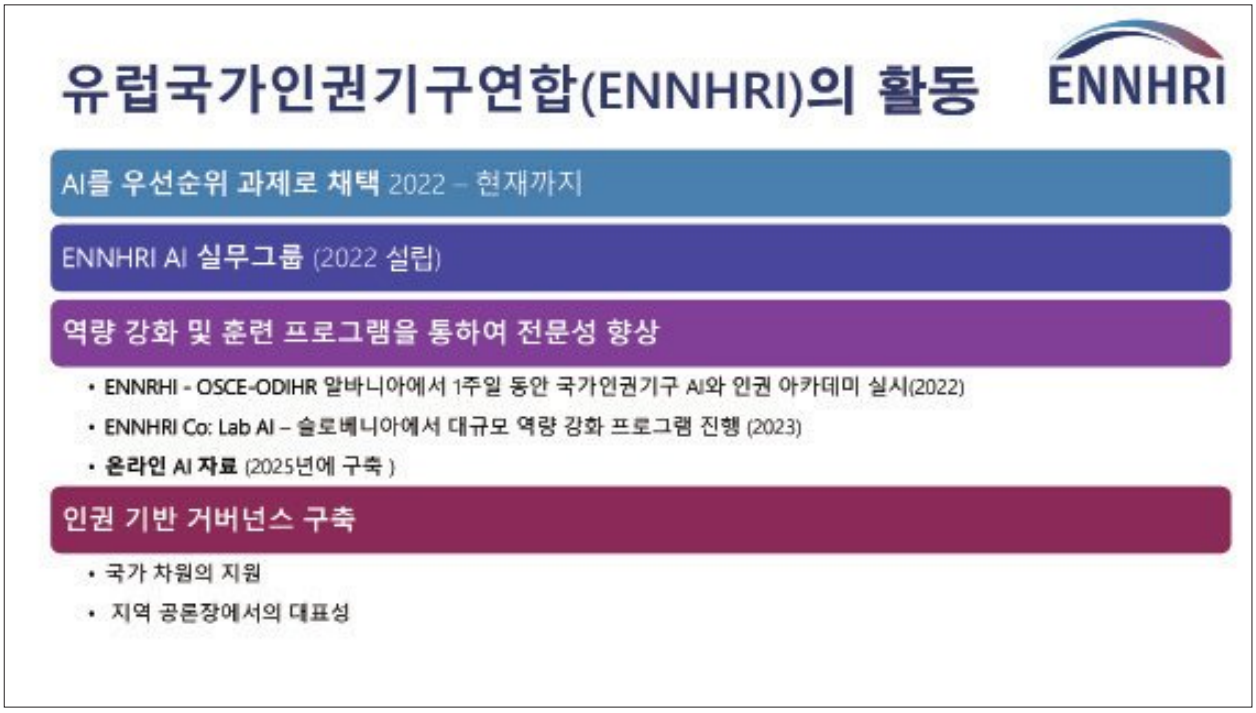


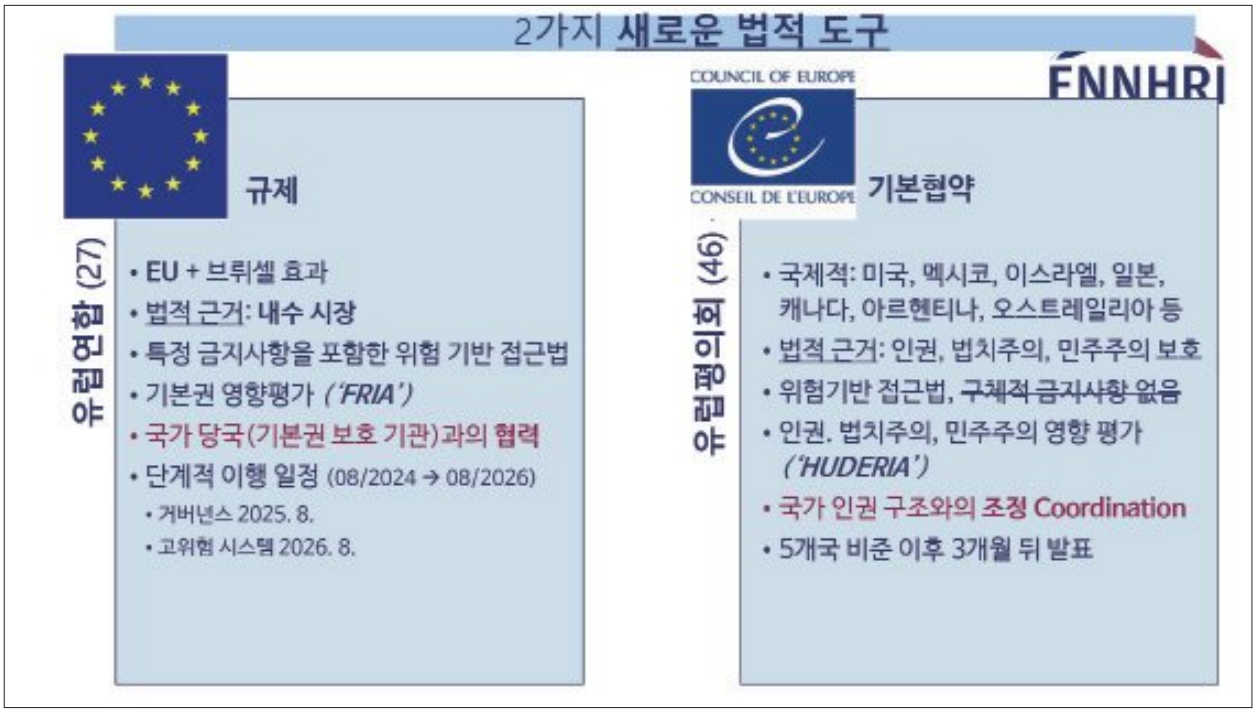
정보의 불균형
지식 격차

권력의 불균형
기준 격차

유럽의 국가인권기구(NHRIs)와 인공지능

- 인공지능과 신기술은 모든 인권에 영향을 미침
  - 광범위한 임무를 지닌 국가인권기구는 AI의 인권적 영향을 다루어야 함.
- 유럽의 국가인권기구들은 AI와 인권 주제에 대해 다양한 수준으로 관여하고 있음
  - 많은 국가인권기구들은 효과적인 대응을 위해 역량, 전문성, 자원이 필요하다고 주장함.


유럽국가인권기구연합(ENNHRI)의 활동
ENNHRI
AI를 우선순위 과제로 채택 2022 – 현재까지
ENNHRI AI 실무그룹 (2022 설립)
역량 강화 및 훈련 프로그램을 통하여 전문성 향상
• ENNRHI - OSCE-ODIHR 알바니아에서 1주일 동안 국가인권기구 AI와 인권 아카데미 실시(2022)
• ENNHRI Co: Lab AI – 슬로베니아에서 대규모 역량 강화 프로그램 진행 (2023)
• 온라인 AI 자료 (2025년에 구축)
인권 기반 거버넌스 구축
• 국가 차원의 지원
• 지역 공론장에서의 대표성


2가지 새로운 법적 도구
유럽연합 (27)
규제
• EU + 브뤼셀 효과
• 법적 근거: 내수 시장
• 특정 금지사항을 포함한 위험 기반 접근법
• 기본권 영향평가 ('FRIA')
• 국가 당국(기본권 보호 기관)과의 협력
• 단계적 이행 일정 (08/2024 → 08/2026)
• 거버넌스 2025. 8.
• 고위험 시스템 2026. 8.
COUNCIL OF EUROPE
CONSEIL DE L'EUROPE
ENNHRI
유럽평의회 (46)
기본협약
• 국제적: 미국, 멕시코, 이스라엘, 일본, 캐나다, 아르헨티나, 오스트레일리아 등
• 법적 근거: 인권, 법치주의, 민주주의 보호
• 위험기반 접근법, ~~구체적 금지사항 없음~~
• 인권. 법치주의, 민주주의 영향 평가 ('HUDERIA')
• 국가 인권 구조와의 조정 Coordination
• 5개국 비준 이후 3개월 뒤 발표

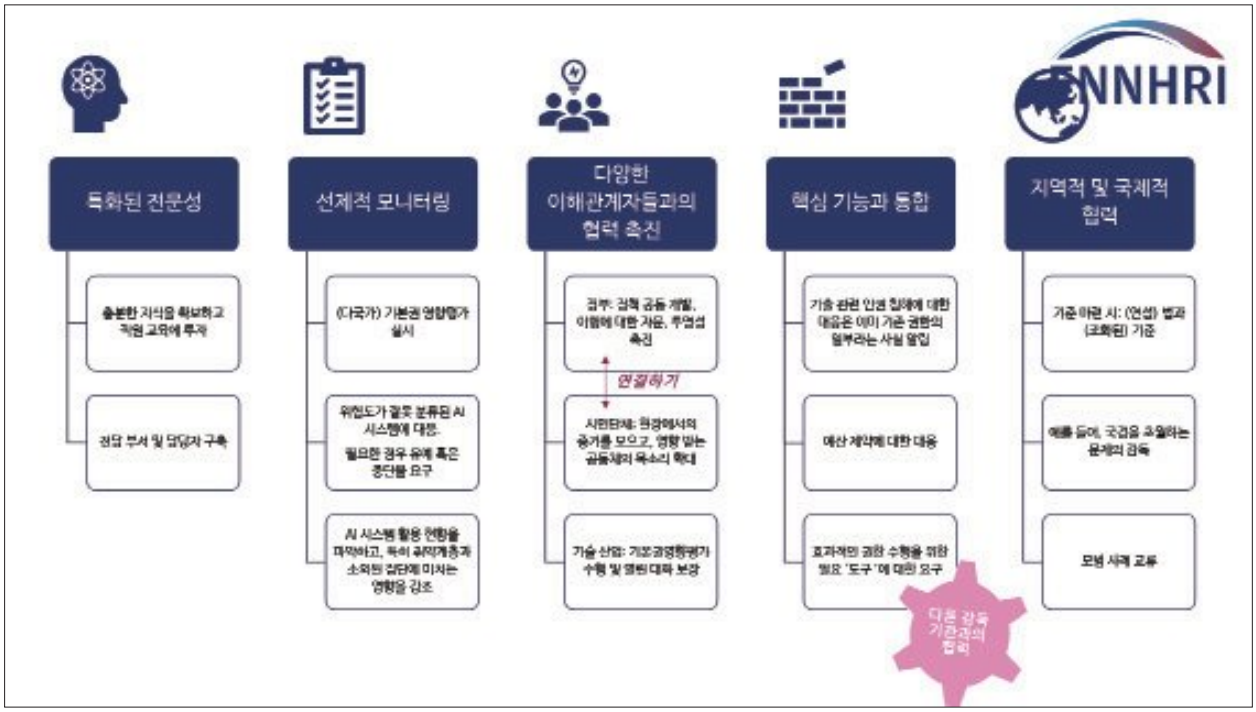
거버넌스에서의 인권 기반 접근 보장 -지역적 참여

ENNHRI

- 유럽 연합 AI 법안
  - ENNHRI 공동 입장
  - ENNHRI-Equinet AI 법안 3자 협의 논의에 대한 성명
- 유럽평의회 AI 위원회 *("CAI")*
  - ENNHRI 유럽평의회 AI위원회에 옵저버 자격이 있고, CDDH-IA에 참여함. (AI와 인권에 대한 핸드북 작성)
  - ENNHRI 공동 입장
  - ENNHRI-Equinet 집행 관련 성명
  - ENNHRI AI 협약 초안에 대한 우려 성명

ENNHRI

국가인권기구들은 어떻게 신기술로 인한 인권 문제를 감시하고, 평가하고, 대응할 수 있을까?



기본권영향평가 들여다보기

기본권 영향평가의 핵심 요건

ENNHRI

1. 기본권 영향 평가 수행자 혹은 평가 팀은 관련한 **인권 전문성을** 보유하거나 관련한 도움을 받아야 함
2. 기본권영향평가의 **1차적 목적은** 인권 침해 **예방**이라는 것을 강조해야 함.
3. **평가의 방법과 기준은** 인권 규범에 부합하여야 함.
4. 기존 국제 기준 및 EU 규제와의 **정책적 일관성을** 촉진해야 함.
5. **실질적 투명성과 효과적 구제를** 보장해야 함.
6. 기본권영향평가는 **집단적 및 사회적 차원의 피해도** 다룰 것을 권고함.
7. 기본권영향평가의 과정에서 **이해관계자의 실질적인 참여를** 보장해야 함.
8. 기본권영향평가는 **반복적인** 과정으로, AI 생애주기 전체를 포괄하는 것으로 이해되어야 함.

전체 내용: https://ennhri.org/wp-content/uploads/2025/04/ENNHRI-statement-on-ensuring-effective-Fundamental-Rights-Impact-Assessments-FRIAs-under-the-EU-AI-Act.pdf

ENNHRI

모범 사례

ENNHRI

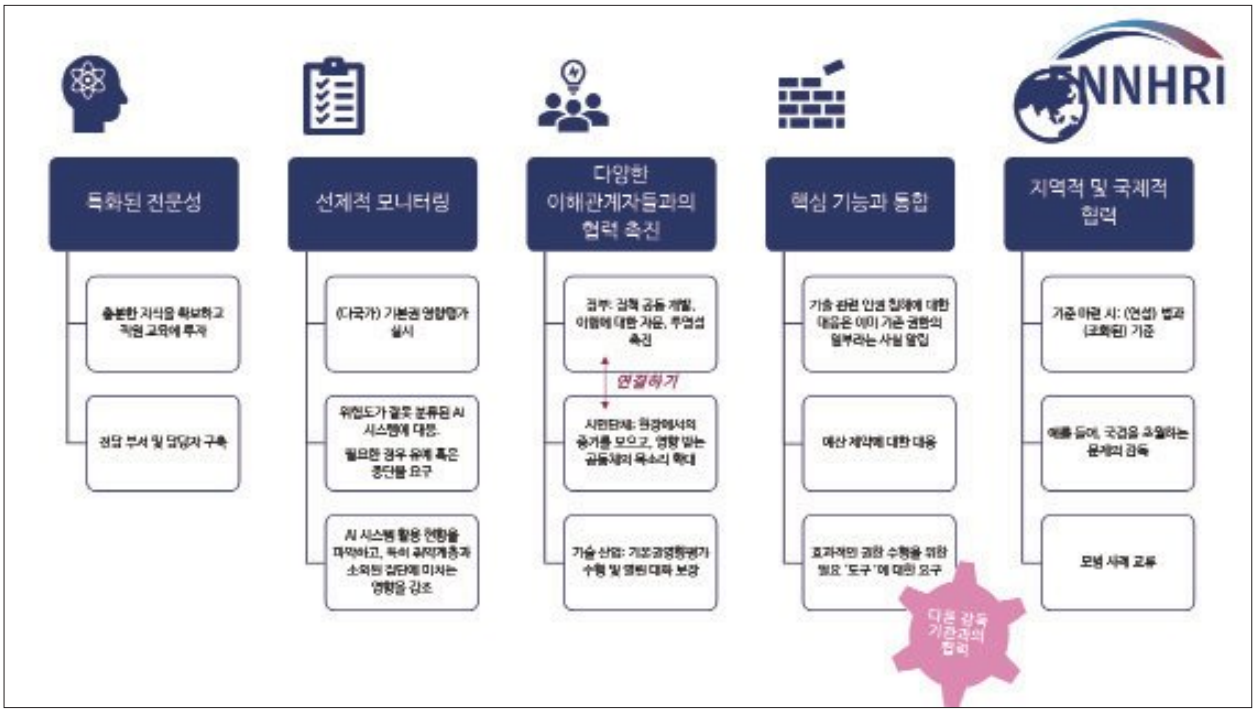
국가인권기구는 어떻게 제도적 권위, 권한, 그리고 운영 역량을 강화할 수 있을까?

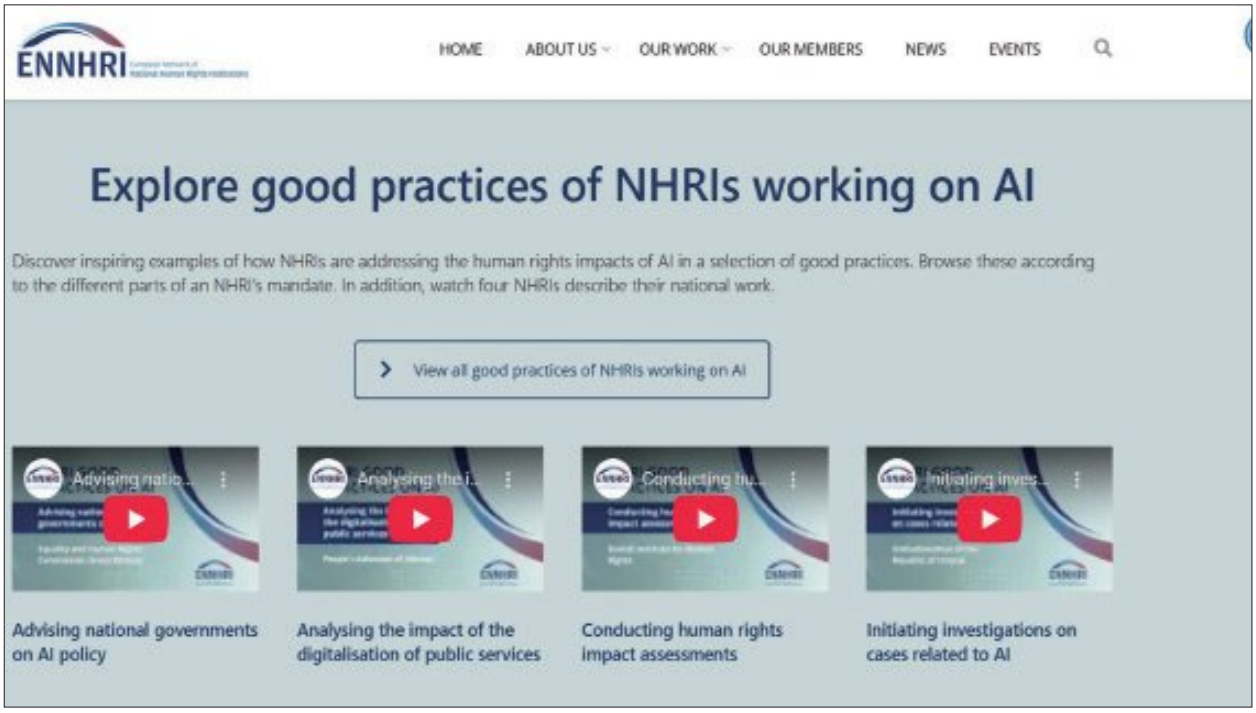

EU
AI
ACT


제도적 권위와 권한 강화하기
모범 사례
ENNHRI
NHRIs
(제77조)
시장감독기관
제77조는 국가인권기구에게 기존 임무를
효과적으로 수행할 수 있는 권한을 부여함.
• 정보접근권 – 제77조 제1항
• 시장감독기관(MSA)의 시험 요청 권한-제77조 제3항
• 기본권 위험이 확인되는 경우
시장감독기관으로부터 정보를 제공 받고
협의함 – art. 79 (2)
현재까지 17개 유럽 국가인권기구가 제77조에 해당하는 기관으로 지정됨.

ENNHRI

어떤 전략을 통해 국가인권기구들은 신기술과 관련된, 변화하는 인권 의제에 대해서 선제적이고 주도적인 역할을 할 수 있을까?



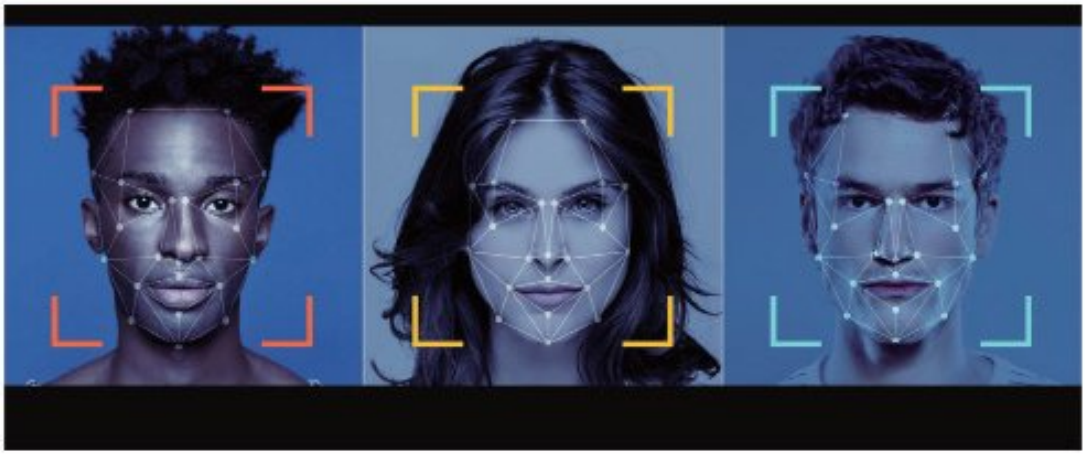

ENNHRI
HOME
ABOUT US
OUR WORK
OUR MEMBERS
NEWS
EVENTS
Explore good practices of NHRIs working on AI
Discover inspiring examples of how NHRIs are addressing the human rights impacts of AI in a selection of good practices. Browse these according to the different parts of an NHRI's mandate. In addition, watch four NHRIs describe their national work.
View all good practices of NHRIs working on AI
Advising national governments on AI policy
Analysing the impact of the digitalisation of public services
Conducting human rights impact assessments
Initiating investigations on cases related to AI


ENNHRI
European Network of
National Human Rights Institutions
감사합니다!
Co-funded by
the European Union

The Role of NHRIs for the New Technology and Human Rights

Nele ROEKENS | Chair of the Working Group on AI, ENNHRI

Human rights challenges



Information asymmetry

Knowledge gap

Power asymmetry

Threshold gap

NHRIs and AI in Europe

- AI and new technologies are affecting all human rights
  - NHRIs – with their broad mandate – to address human rights impacts of AI
- European NHRIs at very different stages regarding their engagement with the topic of AI and human rights
  - Many NHRIs in Europe have reported a need for further capacity, expertise and resources to work effectively on AI

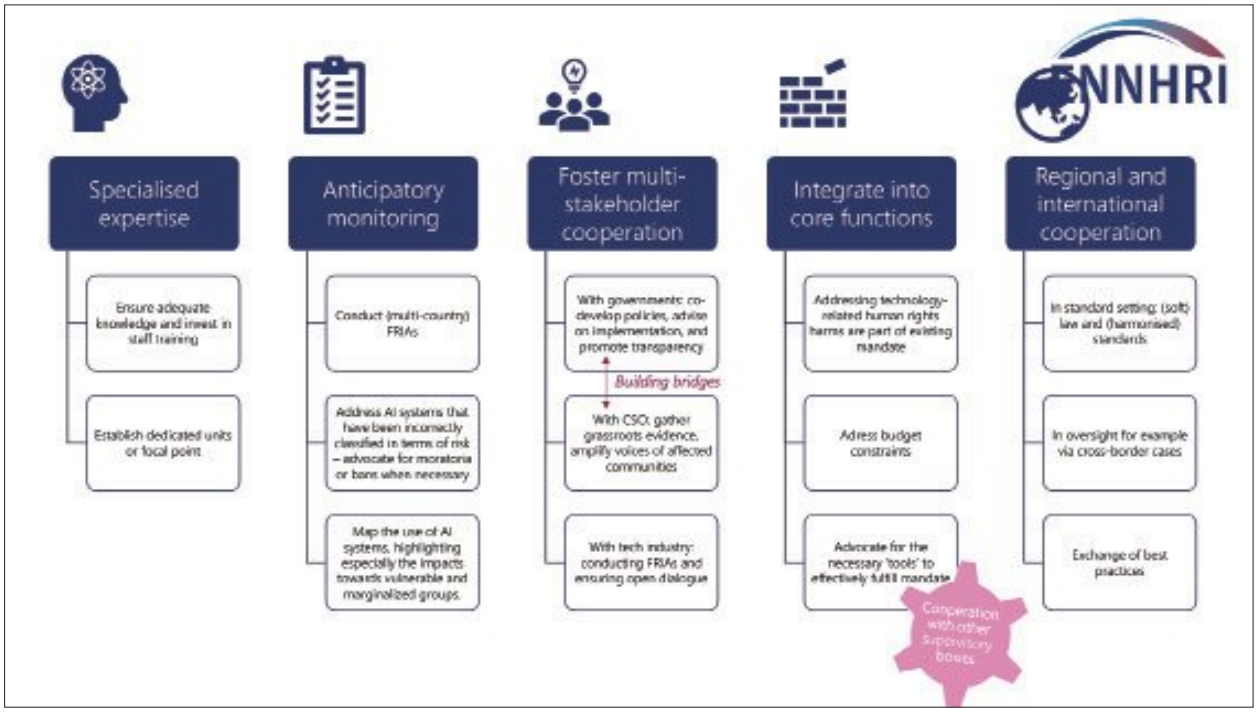

ENNHRI
ENNHRIs activities
AI is a thematic priority 2022 – now
ENNHRI AI Working Group (established 2022)
Developing specialised expertise via capacity building and training programs
• ENNRHI - OSCE-ODIHR weeklong NHRI AI and HR academy in Albania (2022)
• ENNHRI Co: Lab AI - large capacity building event in Slovenia (2023)
• Online AI resource (launched 2025)
Shaping human rights based governance
• Support at national level
• Representation at regional fora


2 NEW LEGAL INSTRUMENTS
ENNHRI
European Union (27)
Regulation
• EU + Brussels effect
• Legal basis: internal market
• Risk based approach including specific prohibitions
• Fundamental Rights impact assesments ('FRIA')
• Cooperation with national authorities protecting FR
• Graduated implementation timeline (08/2024 → 08/2026)
• Governance 08/25
• High-risk systems 08/26
COUNCIL OF EUROPE
CONSEIL DE L'EUROPE
Framework Convention
Council of Europe (46)
• Global: USA, MEX, ISR, JAP, CAN, ARG, AUS etc.
• Legal basis: protection of HR, RoL and democracy
• Risk based approach, no specific prohibitions
• HR, RoL and Democracy Impact Assesment ('HUDERIA')
• Coordination with national HR structures
• Entry into force 3 months after 5 ratifications

Ensuring a human rights-based approach in governance – *regional engagement*

- European Union AI Act
  - ENNHRI Common Position
  - ENNHRI-Equinet statement on AI Act trilogue discussions
- Council of Europe Committee on AI *("CAI")*
  - ENNHRI has **observer status** at CAI and participates in CDDH-IA (drafting handbook on AI and HR)
  - ENNHRI Common Position
  - ENNHRI-Equinet statement on enforcement
  - ENNHRI **Statement of Concern** on draft Convention

How can NHRIs monitor, assess, and respond to human rights issues arising from new technologies?



Zoom in on FRIAs: key requirements for effective FRIAs

ENNHRI

1. Require that individuals or teams performing FRIAs have relevant **HR expertise**, or request relevant assistance.
2. Underline that the **primary purpose** of FRIAs should be the **prevention** of HR violations.
3. Require that the **methodology and benchmarks for assessment** are grounded in HR standards.
4. Promote **policy coherence** with existing international standards and other EU regulations.
5. Require **meaningful transparency** and **effective remedy**.
6. Recommend that FRIAs address **collective and societal-level harms**.
7. Require **meaningful stakeholder engagement** throughout the FRIA process.
8. Ensure the FRIA process is understood as **iterative**, covering the entire AI lifecycle

FULL STATEMENT

ENNHRI

Best practice

ENNHRI

How can NHRIs enhance their institutional authority, mandates, and operational capacity?

Enhancing institutional authority and mandate
Best practice

ENNHRI

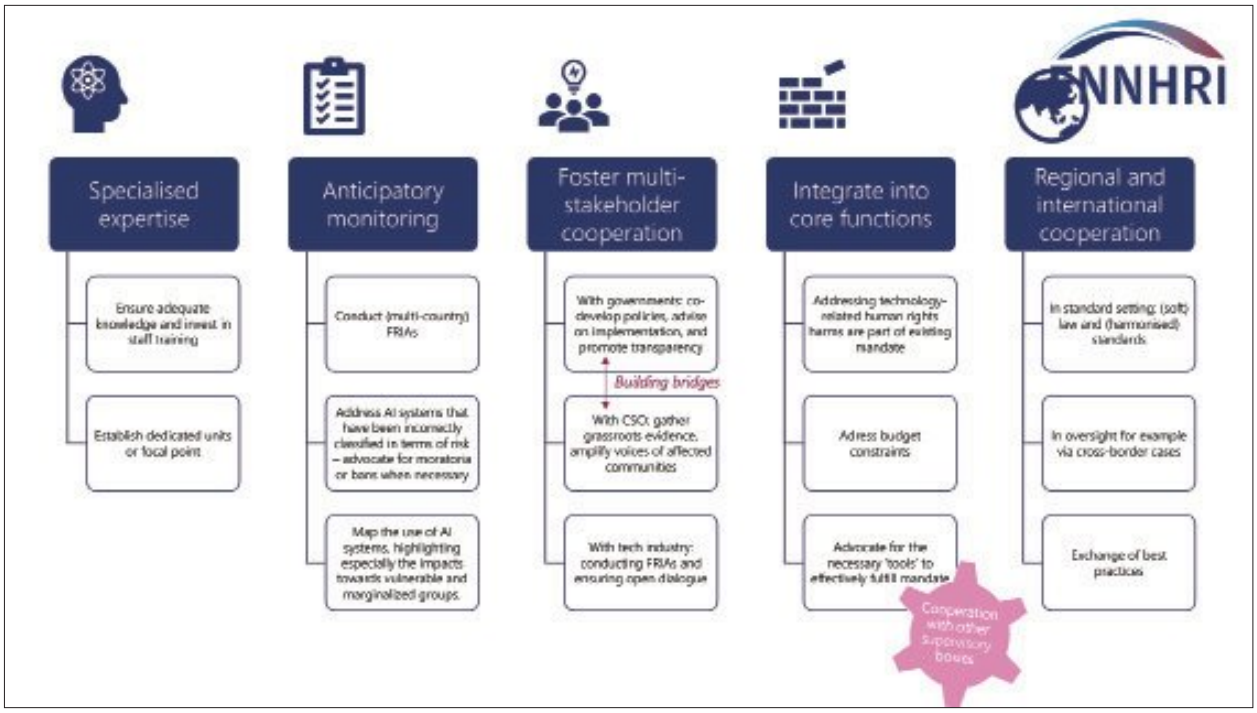
Art. 77 provides NHRIs with the necessary **powers** to effectively **fulfill their existing mandate**

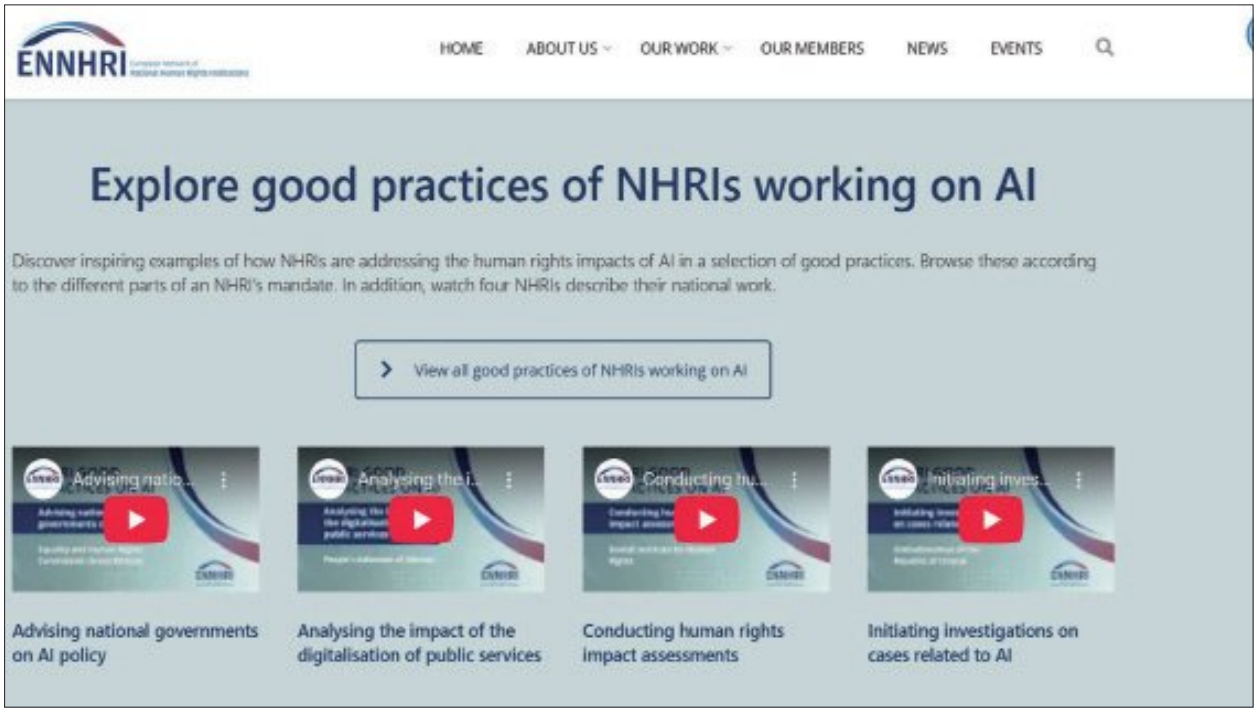
- Access to information – art. 77(1)
- Request testing by MSA – art. 77 (3)
- Being informed *and* consulted by MSA when risks to fundamental rights are identified – art. 79 (2)

So far 17 European NHRIs = article 77 body

ENNHRI

What strategies can NHRIs adopt to take a proactive and leading role in addressing new and evolving human rights agendas related to technology?




ENNHRI
HOME
ABOUT US
OUR WORK
OUR MEMBERS
NEWS
EVENTS
Explore good practices of NHRIs working on AI
Discover inspiring examples of how NHRIs are addressing the human rights impacts of AI in a selection of good practices. Browse these according to the different parts of an NHRI's mandate. In addition, watch four NHRIs describe their national work.
View all good practices of NHRIs working on AI
Advising national governments on AI policy
Analysing the impact of the digitalisation of public services
Conducting human rights impact assessments
Initiating investigations on cases related to AI


ENNHRI
European Network of
National Human Rights Institutions
Questions?
Thank you for your attention!
Co-funded by
the European Union

신기술과 인권: 인공지능의 기회와 도전

New Technology and Human Rights

Opportunities and Challenges of Artificial Intelligence

| 인 쇄 | 2025년 9월

| 발 행 | 2025년 9월

| 발행인 | 안창호(국가인권위원회 위원장)

| 발행처 | 국가인권위원회 국제인권과

| 주 소 | (04551) 서울시 중구 삼일대로 340 나라키움 저동빌딩

| 전 화 | (02) 2125-9883 | F A X | (02) 2125-0918

| Homepage | www.humanrights.go.kr

| 인쇄처 | 디자인모장

| 전 화 | (02) 2278-1990 | F A X | (02) 2278-1992

발간등록번호 11-1620000-100041-01

ISBN 979-11-7214-096-0 93330

발 간 등 록 번 호
11-1620000-100041-01

국가인권위원회

APF

ISBN 979-11-7214-096-0 93330