인공지능 기술 활용에서 편향(성) 최적화 방법론 융합연구

A Convergence Research on Optimization of Biases in AI Application

2020년도 일반공동연구지원사업
(융합연구팀)

경남대학교 융합연구팀
연구책임자: 정원섭

연구 개요

A Convergence Research on Optimization of Biases in AI Application

1

【연구 목적】

연구 목적

최근 인공지능 기술 활용 과정에서 편향의 문제가 가장 중요한 사안으로 부각되고 있는바, 본 연구는 이를 최적화하기 위한 기본 원칙과 가이드라인, 그리고 프로토콜 개발을 목적으로 한다.

세부 목표

본 연구는 이를 통해 인공지능 기술 활용 과정에서 인공지능 기술이 인류 전체의 공동 자산으로 기능할 수 있도록 현재의 다양한 사회적 요구에 부합하면서도 미래 지향적인 규범적 준거를 제시하고, 이를 실증적으로 검증하는 것을 목표로 한다.

융합연구의 필요성

오늘날 인간과 사회의 거울이 되는 인공지능 기술은 새로운 '접경(front)'으로서 학제간 포괄적인 융합 연구를 요구하고 있다. 그것은 바로 인공지능 기술이 인문학, 사회과학, 법학, 자연과학, 공학 그리고 의학 등 현재 인간 지성이 총체적으로 융합하는 새로운 '장소'(locus)이기 때문이다.

【연차별 목표】

/02 편향성의 연원 심층 진단

편향성의 연원에 대한 심층 진단

/04 가이드라인 제시

최적화 기본 원칙을 구체화할 수 있는 가이드라인 제시

/01 인공지능 편향성 조사 분석

인공지능 기술 활용 과정에서 등장한 편향성에 대한 조사 분석

/03 편향성 최적화 기본 원칙 마련

인공지능 편향성을 최적화하기 위한 기본 원칙 마련

/05 프로토콜 개발 및 검증

구체적인 프로토콜을 개발 및 검증

편향성 최적화 연구의 목표

A Convergence Research on

Optimization of Biases in AI Application

부정적 편향 배제

자유와 평등 그리고 우애를 기반으로 한 현대 민주주의 사회에서 이미 공공연히 배제하고 있는 부당한 차별과 억압을 초래할 수 있는 편향들을 배제한다.

기술 활용의 공공성 증진 가능성 모색

인공지능 기술은 인류가 성취한 공동 자산이라는 점에서 소외 계층의 지위를 향상하고 사회적 공공성을 증진하는 방향으로 기술이 활용될 수 있는 가능성을 모색한다.

공공적 개입 가능성 모색

현재 시점에서 인공지능이 맥락에 따라 적절한 작동을 할 수 있도록 공공적 개입의 가능성을 모색한다.

【기대효과】

인공지능 기술 개발에서 인문학의 참여 기회 확장

편향성에 대한 인문학 중심 융합 연구를 통해 인공지능 기술에 대한 사회적 신뢰 증진 및 인공지능 기술 개발 과정에서 인문학의 참여 기회 확장

인공지능 기술의 효율성 및 사회적 공공성 증진

소셜 로봇 및 소셜 알고리즘에 등장할 수 있는 악성 편향을 사전에 차단함으로써 인공지능 기술의 효율성 및 사회적 공공성 증진에 기여

한국 사회의 인문적 소양의 질적 고양

인공지능 및 인간 편향성에 대한 교양 도서 제공

인공지능 바이어스 및 인간 편향성에 대한 훌륭한 교양 도서 제공

편향성 최적화 방법론 융합연구

임상 및 상담 심리 활동에 실질적 기여

바이어스 교정 방안을 통해 임상 심리 및 상담 심리 활동에 실질적 기여

인공지능 기술 관련 정책 방향 제시

(지적) 재산권적 책임, 프라이버시 및 개인정보보호 정책, 고용노동 정책, 노인 및 여성 관련 정책, 소비자보호 정책, 형사사법 정책 방향 제시

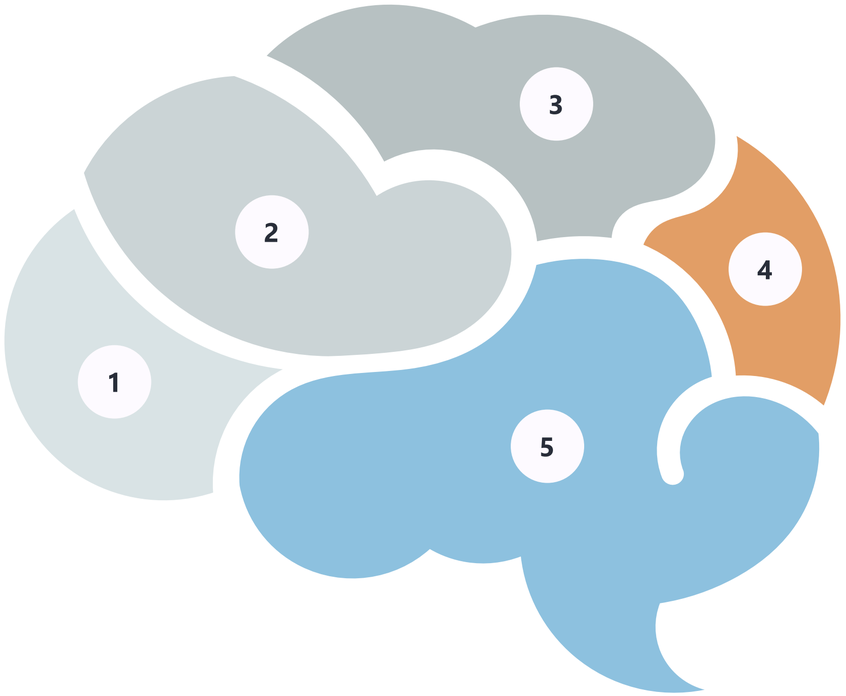
인공지능 활용 관련 입법 및 사법 방향 제시

자율주행차량 사고 및 보험 등 인공지능 기술 활용 과정에서 등장하는 각종 사고와 관련된 민·형사 책임 규명 방향 제시

연구 진행 상황

2

A Convergence Research on Optimization of Biases in AI Application



편향성 최적화 융합연구

연구 내용 및 방법

○ 분석 대상 : 사회적(social) 로봇과 사회적 알고리즘

- 인공지능 활용 과정에서 나타난 편향성에 대한 분석에서 출발
- 소셜 로봇과 사회적 알고리즘에 주목
- 소셜 로봇 및 소셜 알고리즘에 등장할 수 있는 악성 편향을 사전에 차단함으로써 인공지능 기술의 효율성 및 사회적 공공성 증진에 기여

○ 편향(성)의 구분

- 인공지능의 편향 (AI Bias)
  - 알고리즘 및 데이터 편향 (비의도적 집합적 편향)
  - 전문직 편향 (의도적 편향)
- 인간 편향(성) (Human Bias)
  - 근본적-인지적 편향 (Fundamental and Cognitive Bias)
  - 사회적 편향(성) (Social Bias)

○ 악성 편향 인덱스 구축 및 시대적 변화

- 편향 최적화 모듈을 개발하고 이를 바탕으로 차단 목록가이드라인 제시
- 이 과정에서 수집되는 빅데이터에 대한 분석을 통해 시대적 변화 양상 추적

【단계별 연구 목표】

/1단계(1~3년차)

- 공동연구 단행본 출간

가제: 인공지능 편향성에 대한 융합 연구
내용: 인공지능 편향성 관련 문제 현황 진단

- 인공지능 편향성 백서 출간

/2단계(4~5년차)

- 정책 제안

연구진 논문 총서 단행본으로 제시

【단계별 연구 목표】

1년차

인공지능 기술 활용 과정에서
등장한 편향에 대한 조사 분석

3년차

편향성을 최적화하기 위한
기본 모듈 고안

5년차

인공지능 편향성 최적화
프로토콜 개발 및 검증

1단계 성과: 인간 편향 및 인공지능 편향에 대한 융합연구

2단계 목표: 최적화 기본 원칙, 프로토콜, 가이드라인의 개발 및 활용

2년차

편향성의 연원에 대한
융합연구를 통한 심층 진단

4년차

최적화 기본 원칙을 구체화 할 수
있는 원칙 및 가이드라인 제시

【1단계 성과: 인간 편향 및 인공지능 편향에 대한 융합연구】

/01 연구 내용

1. 인공지능 편향 관련 연구 동향 및 문헌 조사
2. 인공지능 편향 예측 알고리즘 조사, 분석
3. 인공지능 편향의 유형과 요인에 관한 연구 수행
4. 인간 편향성 실험연구 개관 및 정리
5. 인간의 새로운 편향성 등장 가능성 연구
6. 인간 편향성 관련 문학 텍스트 자료 분석
7. 기계에 대한 동양적 사유와 편견
8. 문학 작품에서 드러난 젠더 바이어스 수집
9. 계몽 철학자들의 인간 편향성 문헌 연구
10. 자율주행차량 및 의료분야 인공지능 기술의 근거중심 알고리즘과 편향 요인 분석
11. 사회적 로봇과 알고리즘 기반 인공지능에서 바이어스에 대한 종합적 문제를 발굴

1년차

인공지능 기술 활용 과정에서 등장한 편향에 대한 융합연구

/01 성과 및 방법

1. 관련 기술 확보
2. 모델 학습에 필요한 데이터 구축
3. 딥러닝 모델 습득
4. 편향 관련 학습 데이터 구축
5. 편향 관련 기존 연구 조사 및 관련 개념 용어 정의
6. 자율주행차량 및 의료분야 인공지능 기술 설계 방법론 조사 및 근거자료 사례 분석
7. 국내외 소셜 알고리즘 및 소셜 로봇 조사 편향성 분석

【1단계 성과: 인간 편향 및 인공지능 편향에 대한 융합연구】

/02 연구 내용

1. 문제점 분석을 위한 데이터 요인 및 인간 요인으로의 편향성 분류
2. 인간의 편향성과 인공지능 편향의 차이 비교 연구
3. 인간의 편향성과 인공지능 편향을 유발하는 원인에 대한 연구
4. 양자를 비교할 수 있는 틀framework 개발
5. 젠더 바이어스에 관한 정신분석학적 논의와 사회학적 논의의 재구성
6. 기술의 설계와 구현 과정에 다양한 편향성이 중요한 영향을 미친 경우에 대한 성찰 – 역사적 사례 연구와 사회학적 분석을 결합하여 편향성에 대한 케이스 스터디 축적
7. 계몽과 휴머니즘에 대한 찬반 논쟁을 편향성의 견지에서 재구성
8. 자율주행 차량 및 의료분야 인공지능 기술에서 편향성의 소인 도출 및 이에 대한 심층 분석 및 문제점 파악

/02 성과 및 방법

1. 학습 데이터의 word embedding
2. Neural Network model을 이용한 편향 예측 모델 구현
3. 구축된 학습 데이터의 벡터화 모델 구현
4. 편향을 예측하는 딥러닝 모델 프로토타입 설계
5. 학습 데이터의 벡터화 모델 프로토타입

【1단계 성과: 인간 편향 및 인공지능 편향에 대한 융합연구】

/03 연구 내용

1. 편향성 예측 모델의 결과를 시각화
2. 고안된 방법론을 구현할 알고리즘 모듈 개발
3. 사회적 알고리즘 활용 과정에서 등장할 쟁점 발굴 및 정책 방향 제시
4. 인간 편향과 인공지능 편향 상호 비교
5. 분류된 편향이 인공지능 학습에 미치는 영향에 대한 현장 적용 연구
6. 이미 발견된 인간 편향성의 교정 방식에 대한 연구
7. 이전에 발견되지 않았던 인간의 새로운 편향성 출현 가능성 및 이에 대한 교정방법을 연구
8. 인공지능에서 실현 가능한 편향성 교정 모델의 가설 수립
9. 편향성의 기준 변화와 상식을 가진 인공지능 가능성 탐색
10. 연구자와 사용자에 내재된 편향성을 투명하게 드러낸 후 사회적 합의과정을 통하여 최적화시킬 수 있는 프로세스 도출
11. 자율주행차량 및 의료분야 인공지능 기술 활용을 위한 편향 및 요인별 단계별 극복방안 가설을 수립하고 전략을 제시함

3년차

편향성을 최적화하기 위한 기본 모듈 고안

/03 성과 및 방법

1. 인공지능 알고리즘 편향을 예측하는 딥 러닝 모델의 구현
2. 모델 학습 데이터의 확장 구축
3. 통합 모델의 성능 평가 및 개선점 도출
4. 인간 편향 교정 가설 및 모델 제시
5. 자율주행차량 및 의료분야 인공지능 기술 활용을 위한 편향성 및 요인별 단계별 극복방안 가설 수립 및 전략 제시

【2단계 목표: 최적화 기본 원칙, 프로토콜, 가이드라인의 개발 및 활용】

연구 내용

1. 인간 편향성의 유형과 정도를 측정할 수 있는 측정도구의 개발
2. 알고리즘 바이어스 통제를 위한 조건들 제시
3. 교정 모델에 따른 학습 데이터와 알고리즘 실험
4. 인공지능 알고리즘의 형성과정에서 예상 사용자의 피드백을 추가 조사하여, 프로토콜, 가이드라인 반영 메커니즘 구상
5. 체계이론의 적응 및 최적화 논의 및 역설 개념
6. 자율주행차량 및 의료분야 인공지능 기술의 활용을 위한 편향 극복 가이드라인 제시
7. 개발된 알고리즘 모델 성능 측정; 발굴한 법적·법정책적 쟁점 기초 연구

목표 및 전략

1. 인간의 편향성의 유형과 정도 측정도구
   a. 편향 예측 모델 및 word embedding 성능 개선
   b. 모델 최적화 및 다양한 데이터 측정을 통한 성능 개선
   c. 최적화된 편향 예측 모델 완성
2. 인공지능 알고리즘의 형성과정에서 예상 사용자의 피드백을 프로토콜, 가이드라인에 반영시키는 메커니즘
3. 자율주행차량 및 의료분야 인공지능의 활용을 위한 편향 극복 가이드라인

【2단계 목표: 최적화 기본 원칙, 프로토콜, 가이드라인의 개발 및 활용】

연구 내용

1. 개발된 알고리즘 모델 및 프로토콜 시험 및 최적화; 발굴한 법적·법정책적 쟁점 응용 연구
2. 편향성에 관한 통섭(consilience) 이론 재검토
3. 제시된 대안에 대한 심리적, 사회적, 공학적 검증
4. 개인에게 적합화된 편향성 교정 방식 프로그램 개발
   - in-class 교육을 위한 프로그램을 개발, 효과성 연구
   - 인공지능 편향과 인간 편향의 차이를 이해하고 극복할 수 있는 교정프로그램 개발
5. 다양한 영역별 소셜 알고리즘 실험과 검증 후 모델의 확립과 이론적 반성
6. 데이터와 프로토콜, 가이드라인을 통한 알고리즘 변형 실험 – 인간의 편향성과 인공지능 편향에 대한 다학제적 검토
7. 역설과 역설 전개의 관점에서 인공지능 편향성 및 최적화 개념 도출
8. 자율주행 차량 및 의료분야 인공지능 기술 편향성 극복 가이드라인에 대한 논리적 검증

목표 및 전략

1. 편향 예측 모델의 테스트 및 활용
2. 제시된 대안에 대한 심리적, 사회적, 공학적 검증
3. 교육을 통한 교정 프로그램과 internet이나 app기반의 교정프로그램의 개발
4. 데이터와 프로토콜, 가이드라인을 통한 알고리즘 변형 실험 – 인간의 편향과 인공지능 편향에 대한 다학제적 검토

연구 실적

A Convergence Research on Optimization of Biases in AI Application

3

【연구 실적】

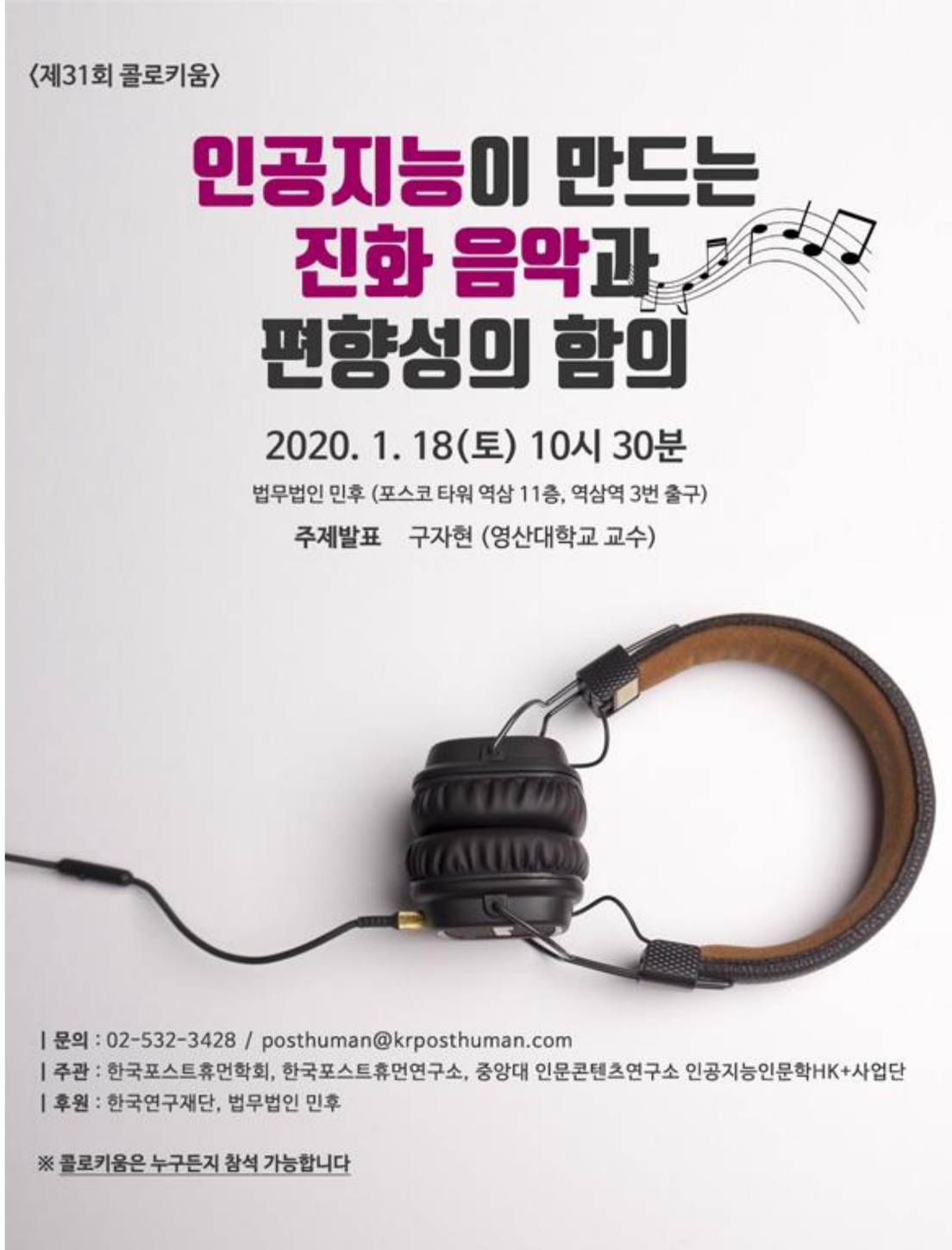
콜로키움

인공지능이 만드는 진화 음악과 편향성의 함의

발표: **구자현** (영산대학교)

일시: 2020년 1월 18일 토요일 오전 10시 30분

장소: 법무법인 민후 (포스코 타워 역삼 11층)



【연구 실적】

학술대회

인간의 편향과 인공지능의 편향

일시: 2020년 7월 17일 금요일 오후 1시

장소: 유튜브 “한국포스트휴먼학회” 실시간 생중계

인공지능 편향성 최적화와 사회적 공공성 증진
기조연설: 정원섭(경남대)

인간의 편향에 맞선 계몽의 역설과 인공지능의 편향
발표: 정성훈(인천대) | 논평: 현영종(서울대)

인공지능의 차별 완화와 공정성 제고
발표: 김건우(광주과학기술원) | 논평: 정채연(포항공과대)

뉴스 기사의 정치적 편향성 탐지 기법
발표: 강승식(국민대) | 논평: 박충식(유원대)

보건의료에서의 AI와 Bias
발표: 장윤정(국립암센터) | 논평: 하대청(광주과학기술원)

인공지능 편향의 비즈니스적 함의와 회피방안
발표: 이성웅(한국IBM) | 논평: 신영택(EA KOREA)

제1차 인공지능 편향성 최적화 학술대회

인간의 편향과 인공지능의 편향

Human Bias and AI Bias

2020년 7월 17일(금) 13:00-17:40 유튜브 채널 "한국포스트휴먼학회"

- **일 시 :** 2020년 7월 17일(금) 13:00-17:40
- **참여방법 :** http://www.krposthuman.com, 유튜브 채널 "한국포스트휴먼학회"

제1차 인공지능 편향성 최적화 학술대회 조직위원회

조직위원장 정 성 훈 (한국포스트휴먼학회 연구이사)

조직위원 백 종 현 (한국포스트휴먼연구소장)
강 진 호 (서울대 철학사상연구소장)
전 영 록 (경남대 교양교육연구소장)
윤 혜 경 (동의대 디그니타스교양교육연구소장)
정 원 섭 (한국포스트휴먼학회장 / 경남대 인공지능 편향성 최적화 연구단장)

공동주최 한국포스트휴먼학회
한국포스트휴먼연구소
경남대학교 교양교육연구소
서울대학교 철학사상연구소
동의대학교 디그니타스교양교육연구소
경남대학교 인공지능 편향성 최적화 연구단

후 원 NRF 한국연구재단
IBM 한국IBM
심경문화재단 철학문화연구소 philcultvre.com

프로그램

제 1 부 13:00 - 15:20

- 사 회 : **박 진** (동의대 철학상담·심리학과)
- 개 회 사 : **정 원 섭** (경남대 자유전공학부)
- 인 사 말 : **강 진 호** (서울대 철학사상연구소장)
- 인 사 말 : **윤 혜 경** (동의대 디그니타스교양교육연구소장)

- 기조발표 (13:10 - 13:30)
  인공지능 편향성 최적화와 사회적 공공성 증진
  발 표 : **정 원 섭** (경남대 자유전공학부)

- 주제발표 1 (13:30 - 14:10)
  인간의 편향에 맞선 계몽의 역설과 인공지능의 편향
  발 표 : **정 성 훈** (인천대 인천학연구원)
  논 평 : **현 영 종** (서울대 철학과)

- 주제발표 2 (14:10 - 14:50)
  인공지능의 차별 완화와 공정성 제고
  발 표 : **김 건 우** (광주과학기술원 기초교육학부)
  논 평 : **정 채 연** (포항공과대 인문사회학부)

- 1차 토론 (14:50 - 15:10)
- 휴 식 (15:10 - 15:20)

제 2 부 15:20 - 17:40

- 사 회 : **김 용 하** (동의대 문화인문교양학부)

- 주제발표 3 (15:20 - 16:00)
  뉴스 기사의 정치적 편향성 탐지 기법
  발 표 : **강 승 식** (국민대 소프트웨어융합대학)
  논 평 : **박 충 식** (유원대 스마트IT학과)

- 주제발표 4 (16:00 - 16:40)
  보건의료에서의 AI와 Bias
  발 표 : **장 운 정** (국립암센터)
  논 평 : **하 대 청** (광주과학기술원 기초교육학부)

- 주제발표 5 (16:40 - 17:20)
  인공지능 편향의 비즈니스적 함의와 회피방안
  발 표 : **이 성 웅** (한국IBM)
  논 평 : **신 영 택** (EA KOREA)

- 2차 토론 (17:20 - 17:40)

【연구 실적】

콜로키움

인공지능 차별과 편향의 연구 주제들

발표: **오요한** (Rensselaer Polytechnic Institute)

일시: 2020년 9월 25일 금요일 오전 10시 (KST)

장소: 온라인 화상회의 (ZOOM)

콜로키움

제6회 세계인문학포럼

「인공지능의 편향성과 공정성」

발표: **정원섭** (경남대학교)

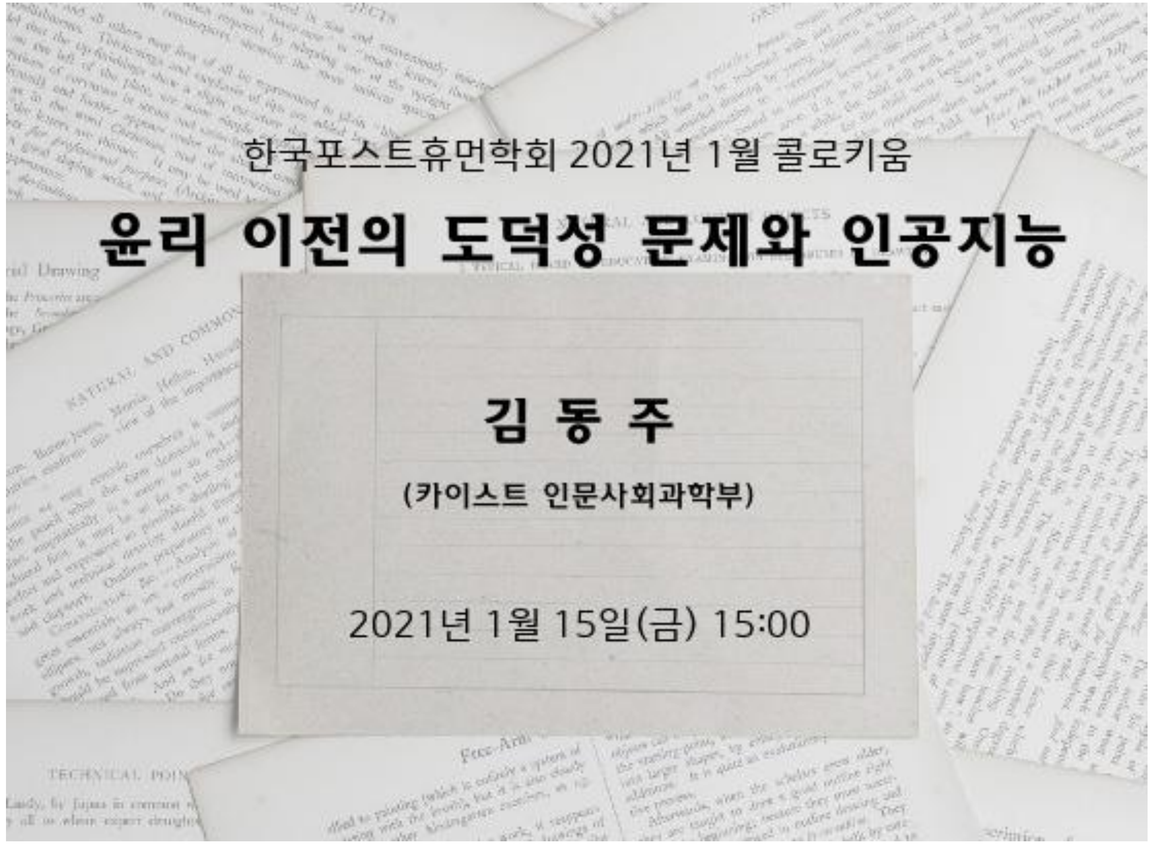
「인공지능의 인간화, 법적 인간화, 신뢰성-인간 중심의 관점과 전망」

발표: **김건우** (광주과학기술원)

일시: 2020년 11월 20일 금요일

장소: 경주화백컨벤션센터

【연구 실적】



콜로키움

윤리 이전의 도덕성 문제와 인공지능

발표: **김동주** (카이스트 인문사회과학부)

일시: 2021년 1월 15일 금요일 오후 3시

장소: 온라인 화상회의 (ZOOM)

【연구 실적】

콜로키움

Enabling Participatory and Procedurally-Fair AI

발표: **이민경** (텍사스 대학교)

일시: 2021년 3월 19일 금요일 오전 10시 (KST)

장소: 온라인 화상회의 (ZOOM)