자율주행 AI의 윤리적 딜레마 분석

MIT Moral Machine 데이터를 활용한 문화권별 판단 기준 비교

지도교수 : 민정혜교수님

202144001 최민석

CONTENTS

브레이크 고장 AI, AI는 누구를 살려야 할까?

자율주행 상용화 시대의 필수 과제: MIT Moral Machine 데이터를 활용한 문화권별 판단 기준 비교 분석

배경 및 필요성

1. 자율주행 기술의 도덕적 공백

레벨 4 상용화 임박, 사고 시 판단 기준(윤리 알고리즘) 부재

2. 문화적 수용성의 차이

획일적 기준 적용 시 사회적 반발 우려

데이터 기반의 문화권별 윤리 인식 분석 필요

프로젝트 목표

1. 글로벌 윤리 데이터 분석

전 세계 52개국 윤리적 선호 경향 시각화 (Save Young 등)

2. 문화권별 핵심 비교 (韓 vs 美)

동양(한국) vs 서양(미국) 가치관 상세 비교 분석

3. 윤리적 현지화(Localization) 제안

국가별 특성 반영 'AI 윤리 옵션(Ethical Knob)' 도입 제시

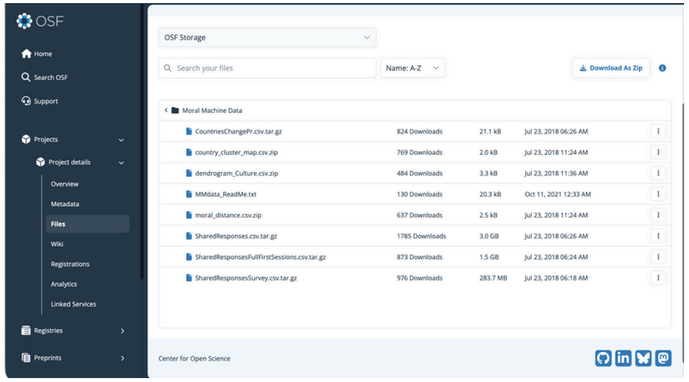
데이터 분석 프로세스

MIT 원본 데이터를 시각화 가능한 형태로 가공하는 3단계 과정

01

MIT 원본 데이터 확보

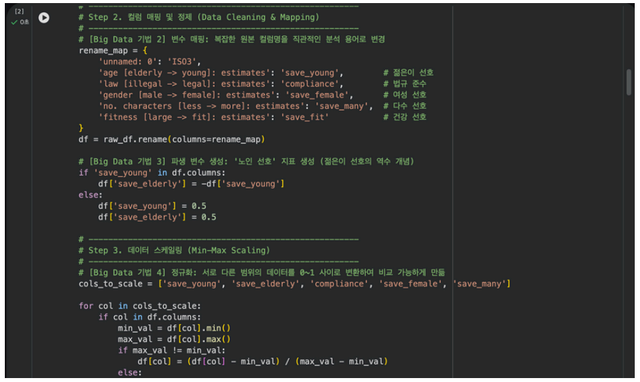
- MIT 원본 데이터 확보 (OSF)
- 신뢰도 높은 주요 52개국 선별



02

데이터 전처리

- 매핑(Mapping): 변수명 변경
- 정규화: 0~1 사이로 값 통일



```
# ------------------------------------------------------------
# Step 2. 컬럼 매핑 및 정제 (Data Cleaning & Mapping)
# ------------------------------------------------------------
# [Big Data 기법 2] 변수 매핑: 복잡한 원본 컬럼명을 직관적인 분석 용어로 변경
rename_map = {
    'unnamed: 0': 'ISO3',
    'age [elderly -> young]: estimates': 'save_young',       # 젊은이 선호
    'law [illegal -> legal]: estimates': 'compliance',       # 법규 준수
    'gender [male -> female]: estimates': 'save_female',     # 여성 선호
    'no. characters [less -> more]: estimates': 'save_many', # 다수 선호
    'fitness [large -> fit]: estimates': 'save_fit'          # 건강 선호
}
df = raw_df.rename(columns=rename_map)

# [Big Data 기법 3] 파생 변수 생성: '노인 선호' 지표 생성 (젊은이 선호의 역수 개념)
if 'save_young' in df.columns:
    df['save_elderly'] = -df['save_young']
else:
    df['save_young'] = 0.5
    df['save_elderly'] = 0.5

# ------------------------------------------------------------
# Step 3. 데이터 스케일링 (Min-Max Scaling)
# ------------------------------------------------------------
# [Big Data 기법 4] 정규화: 서로 다른 범위의 데이터를 0~1 사이로 변환하여 비교 가능하게 만듦
cols_to_scale = ['save_young', 'save_elderly', 'compliance', 'save_female', 'save_many']

for col in cols_to_scale:
    if col in df.columns:
        min_val = df[col].min()
        max_val = df[col].max()
        if max_val != min_val:
            df[col] = (df[col] - min_val) / (max_val - min_val)
        else:
```

03

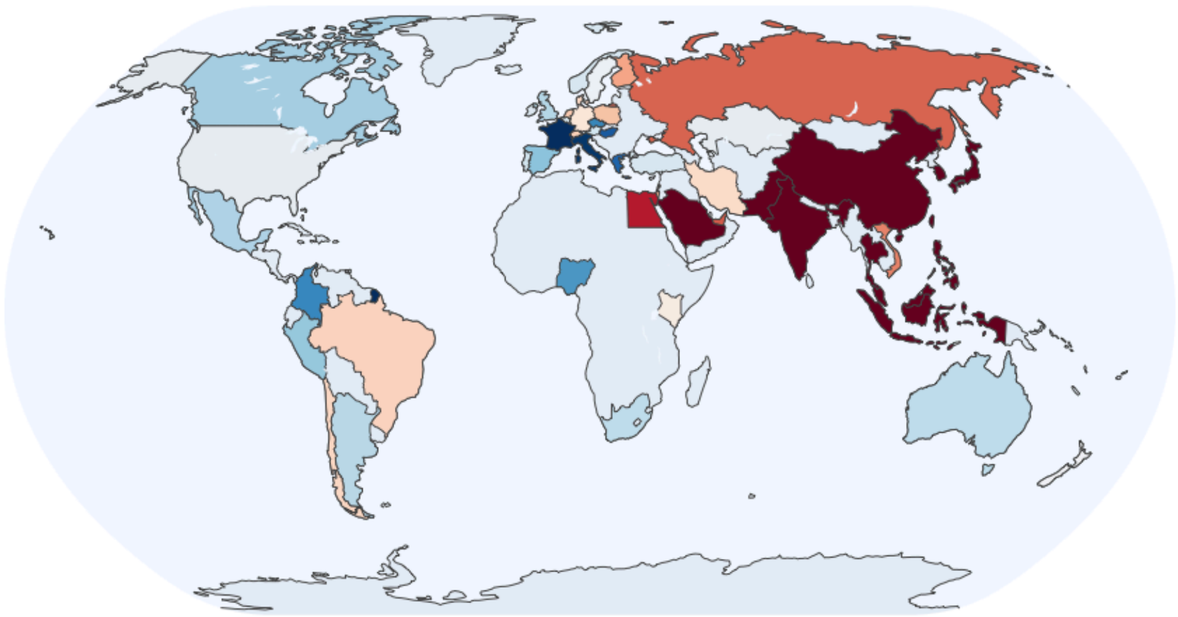
시각화 최적화

- DataFrame 구조화
- Plotly/Matplotlib 포맷 변환

	ISO3	Country	save_young	save_elderly	compliance	save_female	save_many
4	ARE	UAE	0.377813	0.622187	0.462928	0.422907	0.640970
5	ARG	아르헨티나	0.555313	0.444687	0.406574	0.493313	0.545519
7	AUS	호주	0.551243	0.448757	0.349594	0.403130	0.661964
8	AUT	오스트리아	0.536633	0.463367	0.519593	0.340169	0.606082
10	BEL	벨기에	0.529245	0.470755	0.539783	0.486807	0.591526
18	BRA	브라질	0.456388	0.543612	0.501619	0.509345	0.434641
21	CAN	캐나다	0.567462	0.432538	0.368689	0.411436	0.691474
22	CHE	스위스	0.453369	0.546631	0.527146	0.390875	0.545982
23	CHL	칠레	0.458857	0.541143	0.544950	0.534225	0.501693
24	CHN	중국	0.079414	0.920586	0.624688	0.383003	0.212993

[거시 분석] 전 세계 윤리적 성향 분포

'젊은이 선호(Save Young)' 지표를 통한 동·서양 가치관 비교



시각화 개요

- 지표: Save Young (젊은이를 선호하는 정도)
- 범례: 🟦 파란색 (높음) vs 🟥 붉은색 (낮음)

1. 서구권의 특징 (Blue Zone)

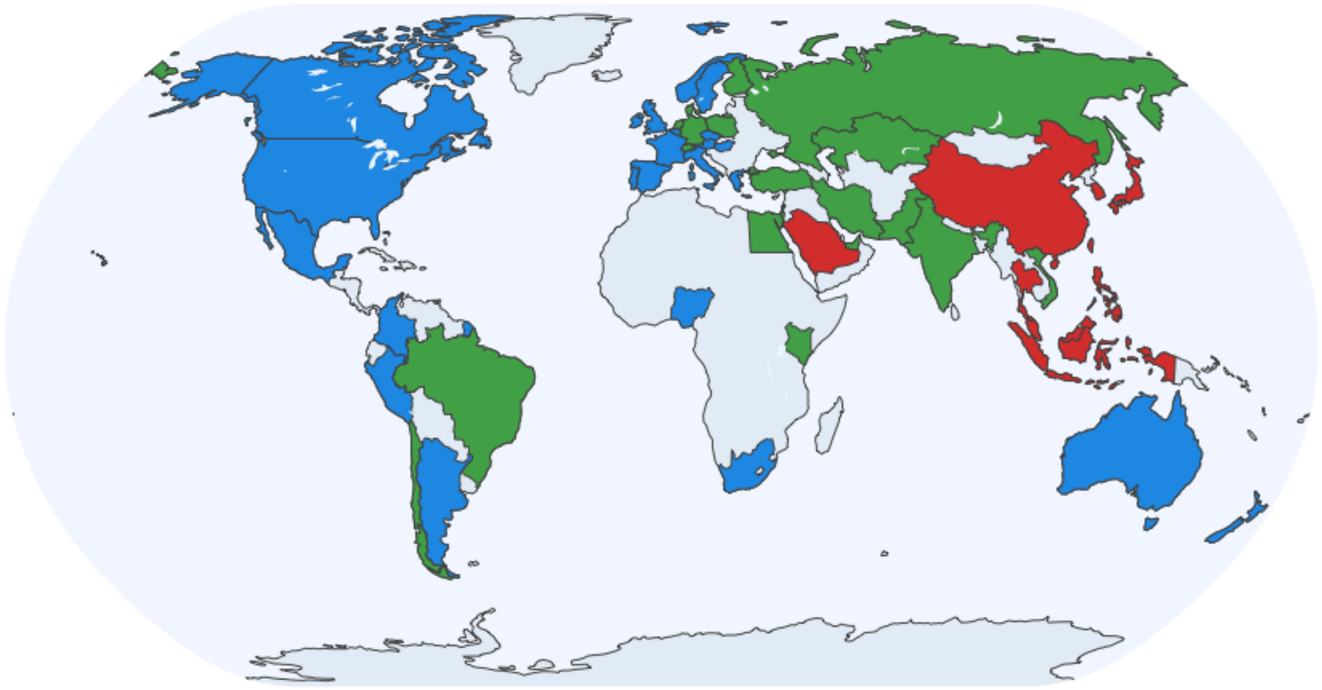
- 미국, 유럽(특히 프랑스, 북유럽) 등은 짙은 파란색으로 나타남.
- **'남은 수명'**과**'미래 가치'**를 중시하여 아이/젊은이를 살리려는 효율성 중심의 사고가 지배적임.

2. 동양권의 특징 (Red Zone)

- 한국, 일본, 중국 등 아시아 국가는 붉은색 계열로 나타남.
- 젊은이 선호보다는 **'노인 공경'**이나 **'사회적 규범'** 등 다른 가치가 의사결정에 더 크게 개입함을 시사함.

[머신러닝] AI가 분류한 세계 윤리 지도

비지도 학습(K-Means Clustering)을 통한 객관적 성향 검증



분석 모델 설계

- 알고리즘: K-Means Clustering (비지도 학습)
- 입력 데이터: 국가명 제외, 오직 5대 윤리 지표 수치만 학습
- 설정: k=3 (3개 그룹으로 분류 요청)

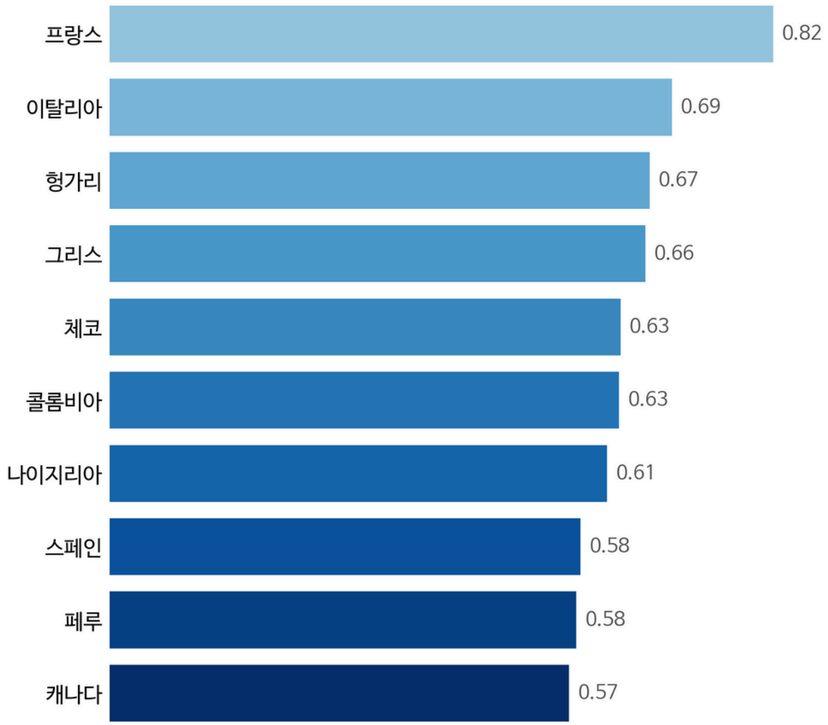
분석 결과

- AI의 판단: 별도의 지역 정보 없이 데이터 패턴만으로 **서구권(파랑/효율)**과 **동양권(빨강/규범)**을 서로 다른 그룹으로 명확히 분류함.
- 의의: 동서양의 윤리적 차이가 주관적 해석이 아닌, 수학적/통계적으로 유의미한 구분임을 머신러닝으로 입증함.

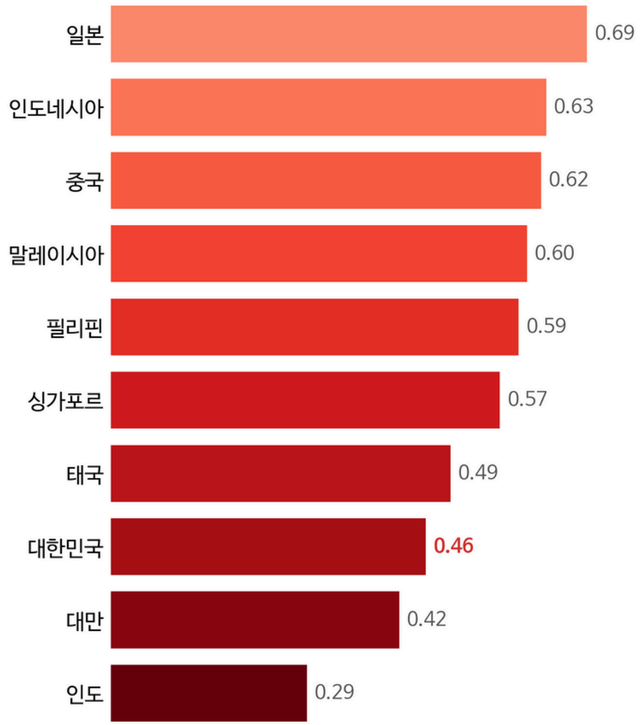
[중간 분석] 왜 `미국`과 `대한민국`인가?

시장 규모와 데이터 특성을 고려한 핵심 타겟 선정

서양 국가 TOP 10: '젊은이 선호'
(북유럽 강세, 미국 상위권)



동양 국가 TOP 10: '법규 준수'
(일본/대한민국 최상위)



글로벌 표준 시장: 미국 (USA)

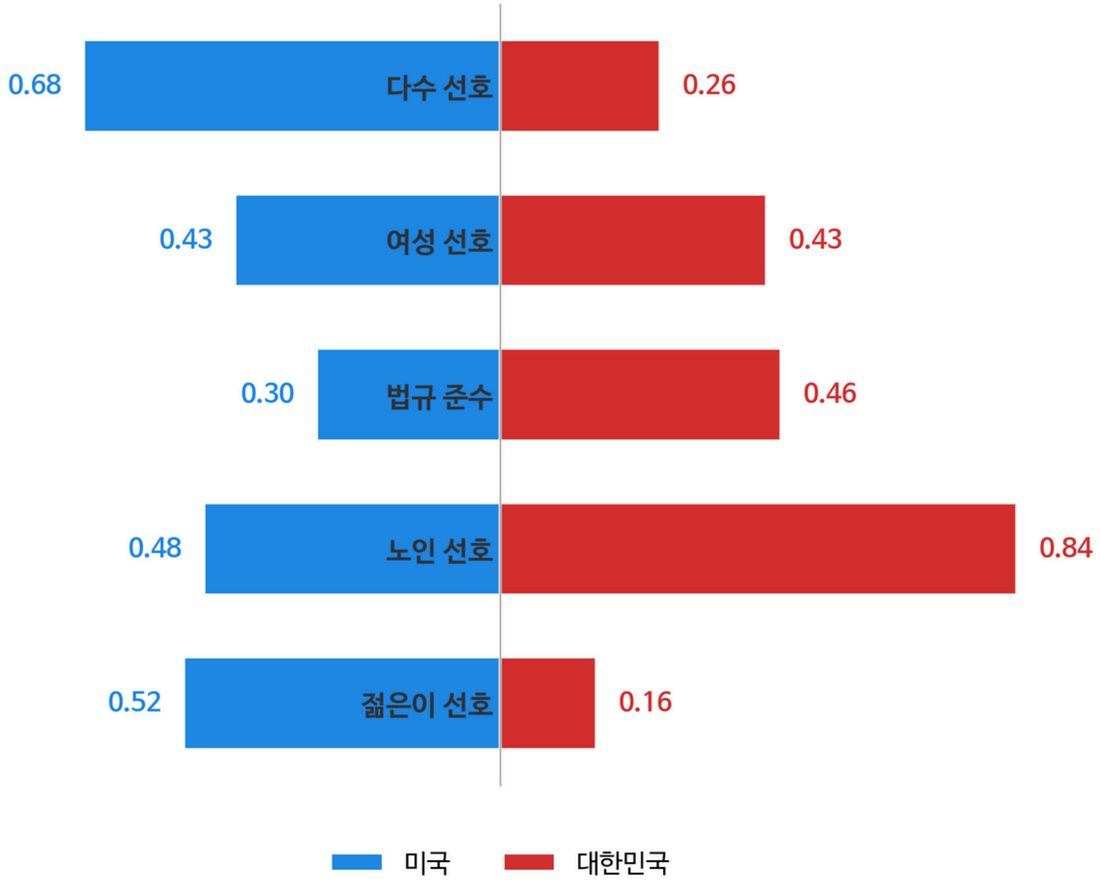
- 선정 이유: 세계 최대 자율주행 시장이자 기술 선도국
- 데이터 팩트: '젊은이 선호' 지표 0.52 (서구권 중상위)
- 해석: 극단적이지 않은 서구권의 보편적인 윤리관을 보여주는 지표로, 글로벌 알고리즘의 기준점으로 적합함

기술 도입 타겟 시장: 대한민국 (KOR)

- 선정 이유: 우리가 개발한 기술이 실제로 주행하게 될 본토 시장
- 데이터 팩트: '법규 준수' 지표 0.46 (동양권 8위)
- 해석: 아시아 국가 중에서도 **사회적 규범**을 중요시하는 **상위 그룹**에 속하므로, 법규 준수 알고리즘이 필수적임.

[미시 분석] 대한민국 vs 미국 핵심 가치 비교

5대 핵심 윤리 지표를 통한 문화적 특성 상세 분석



핵심 가치: 균형 잡힌 효율성

- 수치: 젊은이 선호 0.52 / 법규 준수 0.48 (균형적 분포)
- 특징: 특정 가치에 쏠리지 않고 **효율성(젊은이 구명)**과 질서를 균형 있게 고려하는 합리적 성향을 띔

핵심 가치: 강력한 규범 준수

- 수치: 법규 준수 0.46 (동양권 상위)
- 특징: **"규칙을 어긴 보행자보다 준법 보행자를 우선한다"**는 인식 뚜렷하며, 이는 사고 시 책임 소재를 가리는 중요한 잣대가 됨

[결론] 윤리에 정답은 없지만, `문화적 정답`은 있다

데이터 분석을 통해 도출한 자율주행 AI의 나아갈 길

Global One-Model의 위험성

1. 문화적 불일치 확인: 데이터 분석 결과, 미국(효율 중심)과 한국(규범 중심)의 윤리적 우선순위는 명확히 다름.
2. 도입 시 갈등 우려: 서구권 데이터를 학습한 AI를 한국에 그대로 도입할 경우, **"왜 법을 어긴 사람을 피하지 않는가?"** 와 같은 사회적 비판에 직면할 수 있음.

AI 윤리 옵션 (Ethical Knob) 도입 제안

1. 알고리즘의 이원화: 출시 국가의 문화적 특성(Compliance vs Efficiency)에 맞춰 **기본 윤리 설정값**을 다르게 적용해야 함.
2. 사용자 선택권 보장: 운전자가 직접 윤리 모드(나를 보호 vs 다수를 보호 vs 법규 준수)를 선택할 수 있는 '윤리적 조절 장치(UI)' 구현 필요.

감사합니다.

Q &A

202144001 최민석

자율주행 AI 윤리적 판단 기준 분석 프로젝트