[기말 프로젝트] 자율주행 AI의 윤리적 딜레마 분석

> 부제: MIT Moral Machine 데이터를 활용한 문화권별 판단 기준 비교 분석

---

1. 프로젝트 개요

- **프로젝트명:** 자율주행 AI 윤리 판단 기준의 문화적 차이 분석
- **소속:** 인하공업전문대학 컴퓨터정보공학과 3학년
- **학번/이름:** 202144001 최민석
- **기술 스택:** Python, Pandas, NumPy, Matplotlib, Plotly

📌 연구 배경 및 목적

레벨 4 이상의 완전 자율주행 상용화가 임박했으나, 사고 불가피 상황에서 AI의 판단 기준인 **'윤리적 알고리즘'**은 부재한 상황입니다. 본 프로젝트는 MIT의 대규모 데이터를 분석하여 **'Global One-Model'의 한계**를 지적하고, 데이터에 기반한 **'윤리적 현지화 (Localization)'**의 필요성을 입증하는 것을 목적으로 합니다.

---

2. 데이터 구축 및 전처리 (Data Pipeline)

본 프로젝트는 **빅데이터 분석의 표준 절차(수집-정제-변환-분석)**를 준수하여 수행되었습니다. 특히 대용량 원본 데이터를 처리하기 위해 **Python**의 `Pandas`와 `NumPy` 라이브러리를 적극 활용하였습니다.

Step 1. 데이터 수집 (Data Collection)

- **출처:** MIT Media Lab Open Science Framework (OSF) 저장소
- **수집 방법:** `tar.gz` 형식의 대용량 압축 파일을 Python의 `tarfile` 라이브러리를 활용하여 코드로 직접 해제 및 로드하였습니다. 이를 통해 데이터 파이프라인의 효율성을 확보했습니다.

Step 2. 데이터 전처리 (Data Preprocessing)

- **필터링 (Filtering):** 전 세계 130여 개국 데이터 중, 표본 부족으로 인한 편향(Bias)을 방지하기 위해 데이터 신뢰도가 확보된 **주요 52개국**만을 추출(`isin()`)하여 분석의 정확도를 높였습니다.
- **변수 매핑 (Mapping):** 논문용으로 기재된 추상적인 변수명(예: `utilitarian`, `law` 등)을 데이터 분석가가 직관적으로 이해할 수 있는 변수명(`save_many`, `compliance`)으로 재정의(Renaming)하여 가독성을 개선했습니다.

Step 3. 데이터 변환 및 특성 공학 (Transformation)

- **파생 변수 생성 (Feature Engineering):** 원본 데이터셋에 부재한 **'노인 선호(Save Elderly)'** 속성을 분석하기 위해, 대립 관계에 있는 '젊은이 선호' 속성의 역수 관계로 새로운 특징(Feature)을 생성했습니다.
- **데이터 정규화 (Min-Max Scaling):**

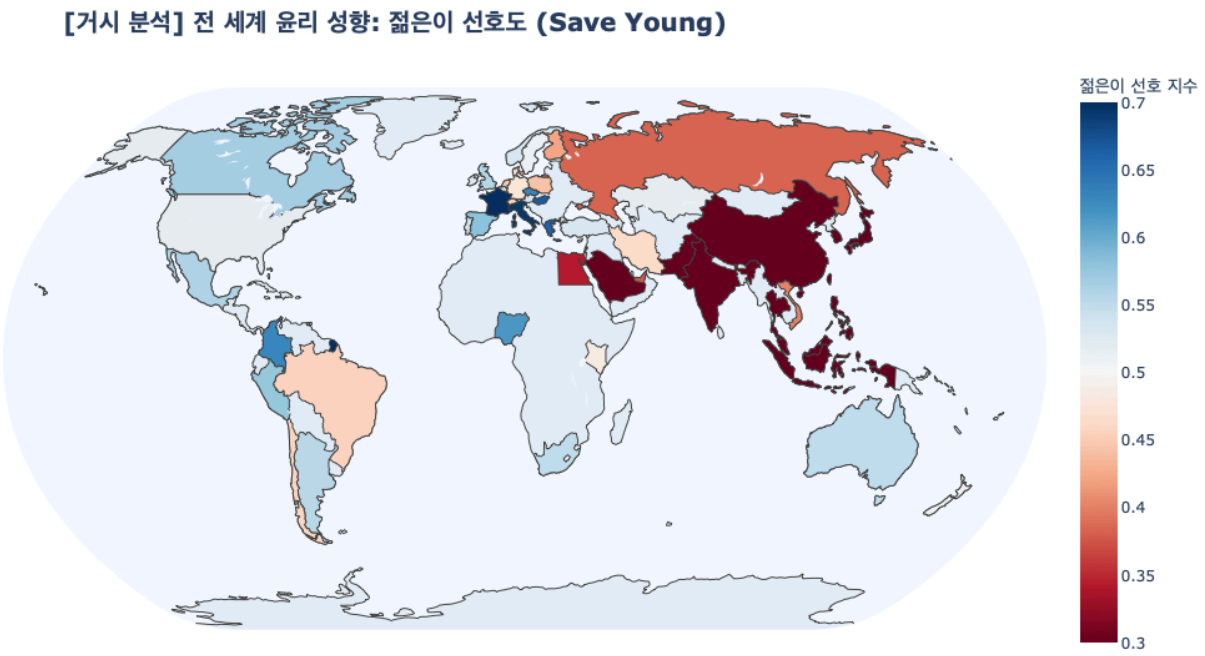
> 국가별/지표별 데이터의 스케일(단위)이 상이하여 직접적인 비교 시 왜곡이 발생할 수 있습니다. 이를 해결하기 위해 모든 수치형 데이터를 0과 1 사이의 값으로 변환하는 최소-최대 정규화를 적용했습니다.

	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

3. 데이터 분석 및 시각화

① [거시 분석] 전 세계 윤리적 성향 지도

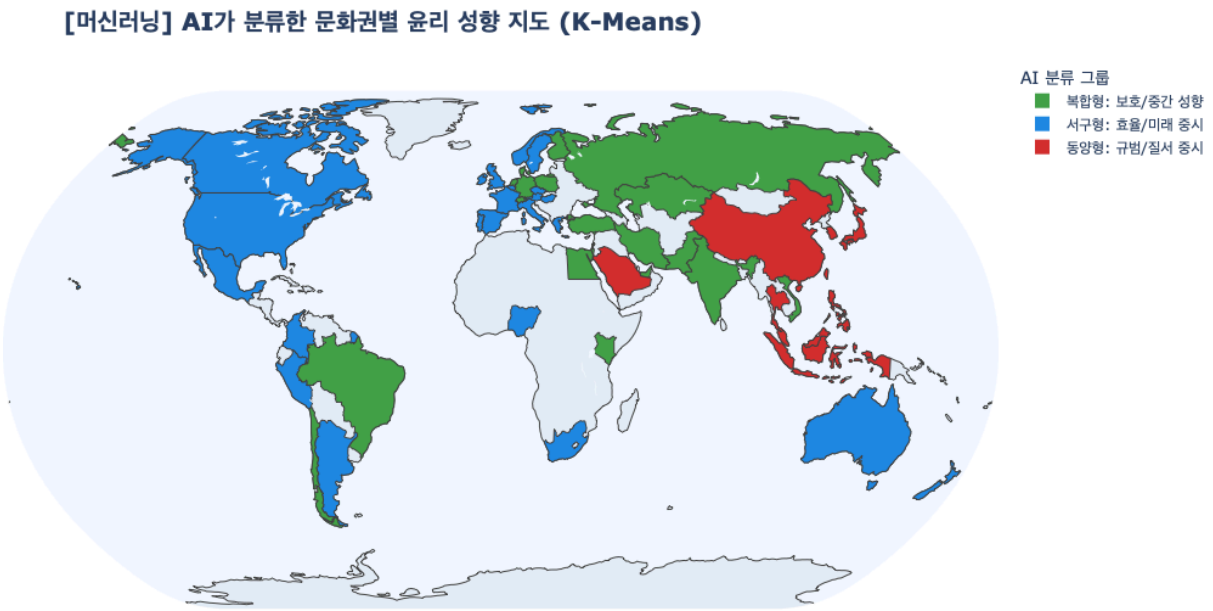
전 세계 50여 개국의 **'젊은이 선호(Save Young)'** 경향을 `Plotly Choropleth`를 활용하여 지도에 매핑했습니다.



> 💡 인사이트
>
> - **서구권 (Blue Zone):** 미국, 유럽 등은 짙은 파란색(High)으로 나타나며, **'남은 수명'**과 **'미래 가치'**를 중시하는 경향이 뚜렷합니다.
> - **동양권 (Red Zone):** 한국, 아시아 등은 붉은색 계열(Low)로 나타나며, 젊은이 선호보다는 **'노인 공경'**이나 **'사회적 규범'** 등 다른 가치가 우선시됨을 확인했습니다.

② [심화 분석] 머신러닝 기반 국가 군집화 (Clustering)

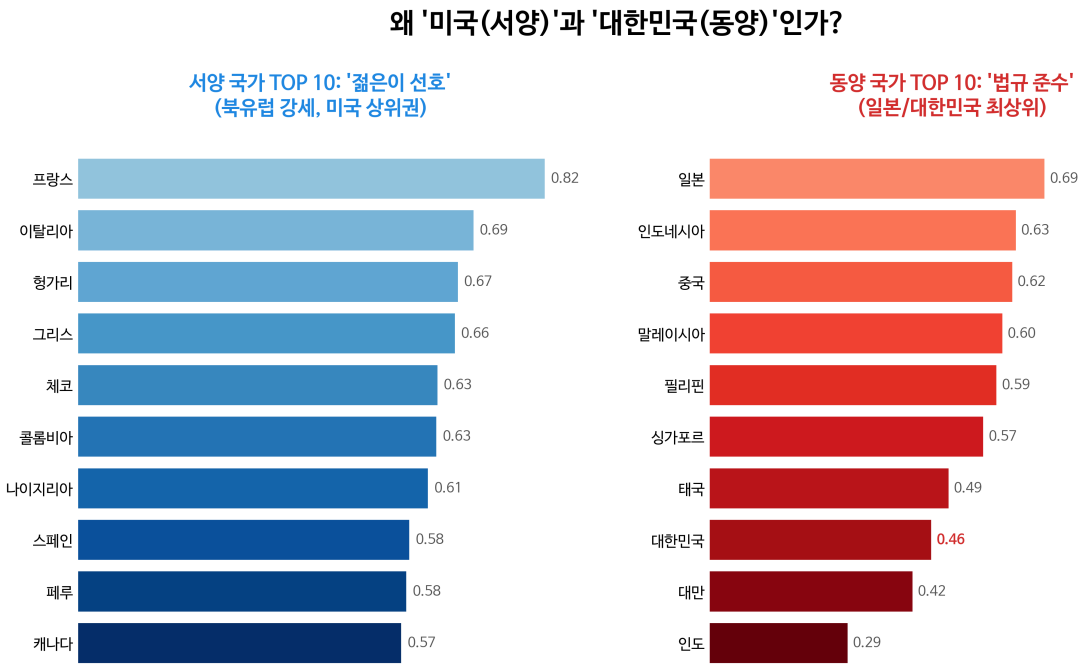
비지도 학습 알고리즘인 **K-Means**를 적용하여 국가별 윤리 성향을 객관적으로 그룹화했습니다.



> 💡 **검증 결과**
>
> - **알고리즘의 판단:** AI는 국가 정보(Label) 없이 오직 데이터 패턴만으로 **서구권(효율 중시)**과 **동양권(규범 중시)**을 수학적으로 다른 그룹으로 분류했습니다.
> - **의의:** 이는 앞서 확인한 문화적 차이가 주관적 해석이 아닌, **통계적으로 유의미한 패턴**임을 입증하는 강력한 근거입니다.

③ [중간 분석] 핵심 타겟 국가 선정

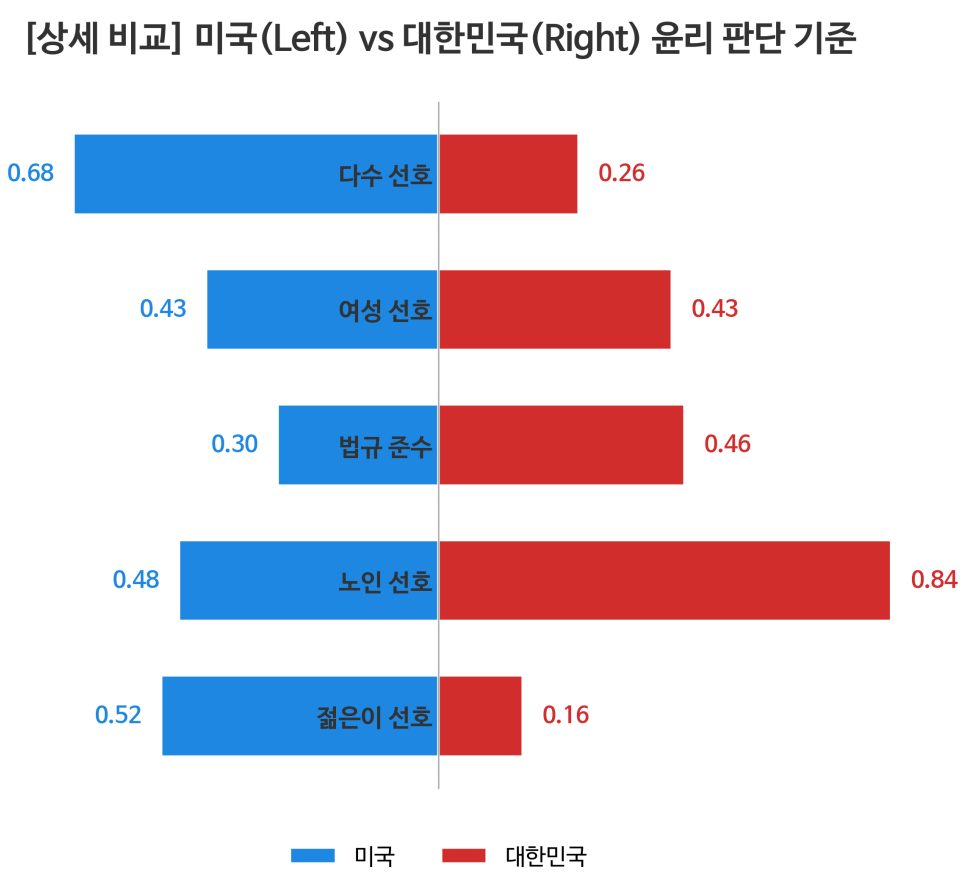
시장 규모와 데이터 특성을 고려하여 동/서양을 대표하는 핵심 국가를 선정했습니다.



> 💡 선정 근거
>
> - 🇺🇸 **미국 (Global Standard):** 세계 최대 자율주행 시장이자, 서구권의 보편적 윤리관(젊은이 선호 0.52)을 대변하는 표준 시장입니다.
> - 🇰🇷 **대한민국 (Target Market):** 기술 도입의 타겟 시장이자, 동양권 중에서도 **'법규 준수(Compliance)' 성향이 최상위권(0.46)**인 특수성을 보여 선정했습니다.

④ [미시 분석] 대한민국 vs 미국 상세 비교

가장 대조적인 두 국가(한국 vs 미국)의 5대 핵심 윤리 지표를 `Butterfly Chart`로 상세 비교했습니다.



> 💡 결론
>
> - **미국 (좌측/파랑):** 위급 상황에서 전체적인 희생을 줄이고 효율성을 추구하는 **공리주의적 성향**이 강합니다.
> - **대한민국 (우측/빨강): "법을 어긴 보행자는 보호받지 못한다"**는 인식이 강하며, 사회적 합의와 규범을 최우선 가치로 둡니다.

4. 최종 결론 및 제언

데이터 분석 결과, 전 세계를 관통하는 단 하나의 **'Global One-Model' 알고리즘은 존재하지 않음**이 증명되었습니다. 서구권 데이터를 학습한 AI를 한국에 그대로 도입할 경우, 사회적 거부감과 법적 분쟁이 발생할 수 있습니다.

따라서 자율주행차 제조사는 다음과 같은 기술적 해결책을 도입해야 합니다.

1. **알고리즘의 현지화 (Localization):** 출시 국가의 문화적 데이터(Compliance vs Efficiency)를 기반으로 기본 윤리 가중치(Weight)를 차별화하여 적용해야 합니다.
2. **윤리적 조절 인터페이스 (Ethical Knob):** 운전자가 자신의 윤리관에 맞춰 주행 모드를 선택할 수 있는 사용자 중심의 옵션 기능을 제공해야 합니다.

📦 Moral-Machine-Analysis
┣ 📂 data # 원본 및 전처리된 데이터셋 (CSV)
┣ 📂 notebooks # 데이터 전처리 및 시각화 소스 코드 (.ipynb)
┣ 📂 results # 시각화 결과 이미지 (Map, Charts)
┗ 📄 README.md # 프로젝트 포트폴리오 문서

---

5. 프로젝트의 한계점 및 향후 과제 (Limitations & Future Work)

① 데이터의 시의성 및 편향성

- **한계:** 본 프로젝트에 사용된 MIT Moral Machine 데이터는 2016~2018년 사이에 주로 수집된 데이터로, 2025년 현재의 윤리적 인식과는 다소 차이가 있을 수 있음. 또한, 온라인 설문 참여자가 디지털 환경에 익숙한 계층으로 편향될 가능성이 있음.
- **개선:** 최신 자율주행 인식 설문 데이터를 추가로 확보하여 시계열(Time-Series) 변화를 분석한다면 더 정교한 결과를 얻을 수 있을 것임.

② 시뮬레이션과 현실의 괴리

- **한계:** '트롤리 딜레마'는 가상의 극단적 상황을 가정한 시뮬레이션 게임이므로, 실제 운전자가 위급 상황에서 내리는 본능적 판단과는 차이가 있을 수 있음.
- **향후 과제:** 실제 교통사고 판례 데이터나 블랙박스 분석 데이터를 활용하여, 사람의 실제 판단 패턴과 설문 결과 간의 상관관계를 분석하는 연구로 확장할 필요가 있음.

6. 참고 문헌 (References)

1. **Data Source:**
    - Awad, E., Dsouza, S., Kim, R. et al. *The Moral Machine experiment*. Nature 563, 59–64 (2018).
    - MIT Media Lab Scalable Cooperation. (2018). *Moral Machine Data*. Open Science Framework. Retrieved from https://osf.io/3hbt7/
2. **Libraries & Tools:**
    - **Python:** Pandas, NumPy, Scikit-learn (K-Means Clustering)
    - **Visualization:** Plotly Express (Choropleth Map), Matplotlib (Butterfly Chart)