자율주행 자동차의 딜레마 시나리오에서 운전자 관점의 도덕 코드

한국인을 대상으로

Driver Moral Codes in Autonomous Vehicles Dilemma Scenarios from Human Driver's Perspective

주 저 자 : 이기쁨 (Lee, Gi-bbeum) 숙명여자대학교 기계시스템학부 석사 후 연구원

공 동 저 자 : 임지민 (Rhim, Jimin) Simon Fraser University, School of Computing Science 박사 후 연구원

공 동 저 자 : 강남우 (Kang, Namwoo) 숙명여자대학교 기계시스템학부 교수

교 신 저 자 : 이지현 (Lee, Ji-hyun) 한국과학기술원 문화기술대학원 교수
jihyunlee@kaist.ac.kr

https://doi.org/10.46248/kidrs.2020.1.125

접수일자 2020. 2. 24. / 심사완료일자 2020. 3. 15. / 게재확정일자 2020. 3. 26.
본 논문은 2017년 한국과학기술대학원 교내연구비를 지원받아 수행되었음.

Abstract

Autonomous vehicles must determine the best action when its unavoidable crashes cause damage to humans. This moral dilemma urges the preparation of a moral system for autonomous vehicles. Motivated by psychological human morality, this study aims to establish drivers' moral codes learned from human reasoning in Autonomous Vehicles moral dilemma scenarios. This study answers the qualitative questions, that is "when" human drivers can get into moral dilemmas in crash situations and "how and why" they make certain moral decisions in such situations. Through the thought experiment of moral dilemma crash scenarios, we organized the moral codes from the perspective of human drivers. Inter-rater reliability of the codes was evaluated as K=0.35, which means fair agreement. The results suggest that human drivers rely on the moral codes of norms, procedures, and actions to make their moral decisions.

Keyword

Autonomous vehicles (자율주행 자동차), AI ethics (인공지능 윤리), Moral reasoning (도덕적 추론)

요약

자율주행 자동차는 피할 수 없는 교통사고로 인해 인명 피해가 일어날 수 있는 상황에서 최선의 행동을 결정해야 한다. 이러한 도덕적 딜레마는 자율주행 자동차에 탑재할 도덕 시스템의 연구를 촉구한다. 도덕심리학적 관점에서 출발하여, 본 연구는 자율주행 자동차의 도덕적 딜레마 시나리오에 대한 사람의 추론을 토대로 운전자의 도덕 코드를 정립하고자 한다. 본 연구는 "언제" 운전자가 교통사고 상황에서 도덕적 딜레마에 빠질 수 있으며, 그 상황에서 "어떻게, 왜" 도덕적 의사결정을 내리는가 라는 질문을 정성적으로 탐구한다. 교통사고의 도덕적 딜레마 시나리오에 대한 사고 실험 결과, 사람의 관점에서 운전자의 도덕 코드를 구조화하였다. 평정자 간 신뢰도는 K=0.35의 일치로 측정되었다. 본 연구의 결과는 운전자가 규범, 절차, 행동적 도덕 코드에 의거하여 도덕적 의사결정을 내림을 시사한다.

목차

1. 서론

1-1. 연구 배경

최근 세계적 기업들의 자율주행 연구는 자율주행 자동차가 실제 무대에 상용화될 날이 머지않았음을 보여준다. 자율주행 자동차가 완전히 도입되면 우리 사회는 교통사고의 감소, 도시 교통 체증 완화, 장애인 및 노약자가 이용할 수 있는 교통수단 제공 등 다양한 이점을 얻을 수 있을 것으로 기대된다.[1)]

그러나 사고를 예측할 수 없는 실제 도로에 자율주행 자동차를 도입하려면 인공지능의 도덕체계에 대한 고민을 피할 수 없다. 교통사고 앞에서 직관적으로 행동하는 인간과 대조적으로, 인공지능은 체계적 의사결정을 내릴 능력이 있기 때문이다.[2)] 이에 "피할 수 없는 교통사고를 앞두고 자율주행 자동차가 누구를 보호해야 하는가" 라는 고민, 즉 자율주행 자동차의 도덕적 딜레마가 발생한다.[3)]

본 연구는 "자율주행 자동차의 도덕론이 확립되지 않은 상황에서, 어떠한 도덕론이 탑재되어야 하는가?" 라는 질문을 탐구하고자 한다. 자율주행 자동차의 도덕론으로서 공공의 피해를 최소화하는 공리주의가 활발히 논의되고 있다.[4)5)6)] 그러나 특정 윤리원칙으로 사회적 합의를 이루기는 쉽지 않다.[7)] 사람들이 제삼자로서 기대하는 윤리원칙과 탑승 당사자로서의 기대 사이에는 간격이 존재한다.[8)] 이러한 간격 사이에서 자율주행 자동차는 비난의 대상이 되거나, 도덕 행위자(moral agent)로 신뢰받지 못하고 희생자에게 책임을 돌릴 수도 있다.[9)] 따라서 자율주행 자동차의 행동이 대중에게 수용되려면, 사람이 공감할 수 있는 도덕 코드에 따라 의사결정을 내릴 필요가 있다. 또한, 도덕 행위자로서 의사결정의 근거를 설명할 수 있어야 한다.

1-2. 연구 목적 및 방법

본 연구는 도덕 행위자인 사람이 자율주행 자동차의 도덕적 딜레마 상황에서 의사결정을 내리는 과정을 이해하는 것을 목적으로 한다. 이를 위하여 본 연구는 1) 실제 교통사고 사례를 바탕으로 자율주행 자동차의 도덕적 딜레마가 일어나는 상황을 고안하고, 2) 도덕적 딜레마 상황에서 운전자의 도덕적 추론에 사용되는 가치요인을 추출한다.

먼저 실제 교통사고 사례 데이터베이스와 조사 보고서를 토대로 운전 경험자들과 심층 인터뷰를 진행한다. 인터뷰를 통하여 운전자의 도덕적 갈등과 요인에 대한 통찰을 얻고, 이를 바탕으로 교통사고 사례 기반의 도덕적 딜레마 시나리오 설계한다. 다음으로, 도덕적 딜레마 시나리오에 대한 의사결정 실험을 진행한다. 이때 사고 구술(think aloud) 방식으로 참여자의 도덕적 추론 과정을 언어로 기록한다. 마지막으로 프로토콜을 분석하여 참여자의 도덕적 추론에 사용된 가치요인을 추출 및 평가한다.

1) Waldrop, M., A World of Driverless Cars, Nature, Vol.518, No.7537, 2015, p.20.
2) Dilich, M. A., Kopernik, D., & Goebelbecker, J., Evaluating driver response to a sudden emergency: Issues of expectancy, emotional arousal and uncertainty, SAE Transactions, 2002, p.238-248.
3) Lin, P., Why Ethics Matters for Autonomous Cars, Autonomous Driving, 2016, p.69-85.
4) Goodall, N. J., Ethical decision making during automated vehicle crashes, Transportation Research Record, Vol.2424, No.1, 2013, p.58-65.
5) Bonnefon, J. F., Shariff, A., & Rahwan, I., The social dilemma of autonomous vehicles, Science, Vol.352, No.6293, 2016, p.1573-1576.
6) Greene, J. D., Our Driverless Dilemma, Science, Vol.352, No.6293, 2016, p.1514-15.
7) Lin P., Op. cit., p.69-85.
8) Bonnefon, J. F. et al., Op. cit., p.1573-1576.
9) Danielson, P., Surprising judgments about robot drivers: Experiments on rising expectations and blaming humans, Etikk i praksis - Nordic Journal of Applied Ethics, Vol.9, No.1, 2015, p.73-86.

2. 이론적 배경

2-1. 교통사고 시 사람의 행동 및 도덕적 의사결정

사람은 운전 중 충돌을 피하기 위해 브레이크를 밟거나, 핸들을 꺾거나, 엑셀을 밟을 수 있다. 그런데 대부분의 운전자는 충돌 위험을 인지하면 단순히 브레이크를 밟거나, 빈 공간으로 핸들을 꺾는 패턴을 보인다.[10][11] 이는 교통사고를 직감한 운전자의 행동이 무의식적인 반사 반응에 가깝다는 것을 보여준다.

대조적으로, 사람이 가치판단 문제를 해석하고 결정하는 과정은 인지적인 추론 활동이다. 이를 도덕적 추론이라고 하는데, 도덕적 추론은 정의(justice)와 걱정(care)의 패러다임으로 해석할 수 있다.[12] 예를 들어, 운전자는 보호받아야 하는 존재에 대한 걱정에 따라 자동차에 태운 어린 아들을 다른 사람보다 우선하기로 결정할 수 있다. 또 다른 운전자는 정의의 윤리에 따라 타인에게 해를 입히는 행동 자체가 비도덕적이라고 판단해 아무런 행동을 취하지 않을 수도 있다.[13]

이와 같이 도덕적 추론은 반사적인 충돌 회피 행동과 달리 인지적인 활동이다. 사람은 교통사고와 같은 급박한 상황에서 도덕적 추론을 수행할 수 없다. 반면 자율주행 자동차는 같은 상황에서도 신호처리에 기반하여 체계적 의사결정이 가능하기에, 자율주행 자동차에게 주어지는 도덕적 책임이 강조된다.

2-2. 자율주행 자동차의 도덕적 의사결정 문제

자율주행 자동차의 출현으로 인해, 트롤리 딜레마(Trolley Dilemma)와 같은 고전적인 윤리 문제가 현실적 의사결정을 요구하게 되었다. 예를 들어, 자율주행 자동차가 보행자나 승객 중 한쪽만 보호할 수 있다면 누구를 보호해야 하는가?[14] 충돌하는 대상이 통학버스라면 자율주행 자동차의 승객이 희생되어야 하는가?[15] 자율주행 자동차로 인해 사고가 발생하면 대중의 도덕적 책임과 비난은 누구에게 향할 것인가?[16] 이러한 질문들은 자율주행 자동차의 상용화에 앞서 도덕 시스템과 정책 마련에 대한 사회적 요구를 보여준다.

인공지능의 도덕적 의사결정은 명확한 윤리원칙을 기반으로 프로그래밍 되는 하향식 방법으로 구현될 수 있다. 자율주행 자동차의 윤리원칙으로서, 사람의 수가 더 많은 쪽을 보호해야 한다는 공리주의적 의견이 지지받는 경향이 있다.[17] 그린(Greene)은 자율주행 자동차가 공공 교통수단이 되면 공리주의적 정책이 정착되리라 전망한다.[18] 그러나 공리주의는 소수의 행복은 보장하지 않으며, 원칙적 체계에 의존하여 인간의 도덕적 가치관이 무시될 수 있다는 위험이 있다.

인공지능의 도덕적 의사결정을 구현하는 또 다른 방법은 상향식 방법으로, 사람이 내린 대량의 의사결정 사례로부터 법칙을 학습하는 것이다. 예를 들어, 모럴 머신(Moral Machine)은 자율주행 자동차의 도덕적 딜레마에 대한 웹 서베이 플랫폼으로, 대량의 의사결정 데이터를 수집하고 대중의 도덕적 기호를 분석한다.[19] 그러나 시나리오가 자율주행 자동차에 대한 삼인칭 시점으로 제한되어 있어 사고 당사자의 입장을 반영하는 데는 한계가 있다. 또한 참여자에게 그들의 의사결정에 대한 "왜"라는 질문을 던지지 않는다. 자율주행 자동차가 스스로의 의사결정을 설명하는 투명한(transparent) 도덕 행위자로 설계되려면, 대량의 사례를 보충하는 정성적 자료가 마련되어야 한다.

3. 연구 방법

3-1. 사고 구술 방법

본 연구는 도덕적 딜레마에 대한 사고 실험(thought experiment)에 사고 구술 방법을 적용하였다. 사고 구술은 참여자가 주어진 일을 해결하면서 생

10) Itoh, M., Lemoine, M. P. P., Robache, F., & Morvan, H., An analysis of driver's avoiding maneuver in a highly emergency situation, SICE Journal of Control, Measurement, and System Integration, Vol.8, No.1, 2015, p.27-33.
11) Dilich. M. A., Op. cit., p.238-248.
12) Wildermuth, C., e Souza, C. A. D. M., & Kozitza, T., Circles of ethics: the impact of proximity on moral reasoning, Journal of business ethics, Vol.140, No.1, 2017, p.17-42.
13) Cushman, F., Young, L., & Hauser, M., The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm, Psychological science, Vol.17, No.12, 2006, p.1082-1089.
14) Bonnefon, J. F. et al., Op. cit., p.1573-1576.
15) Lin, P., Op. cit., p.69-85.
16) Danielson, P., Op. cit., p.73-86.
17) Bonnefon, J. F. et al., Op. cit., p.1573-1576.
18) Greene, J. D., Op. cit., p.1514-15.
19) Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I., The moral machine experiment, Nature, Vol.563, No.7729, 2018, p.59-64.

각하는 것을 모두 소리내어 말하는 구술 보고 방법이다.[20] 이 방법은 문제 해결 과정에서 일어나는 정보의 흐름에 대하여 신뢰할 수 있는 자료를 수집하기 위하여 사용된다.[21] 도덕심리학 연구 분야에서도 인간의 의사결정 중 도덕적 추론을 관찰하기 위한 방법론으로 사용되었다.[22]

3-2. 프로토콜 분석

본 연구에서는 코딩을 통하여 사고 구술 프로토콜을 분석하였다. 코딩은 텍스트에서 중요한 개념을 추출하여 코드로 정의하는 과정이다. 특히 개방 코딩(open coding)은 광범위한 분야의 질적 연구에서 사용되는데, 텍스트에서 찾아낸 개념을 정의하고 개념들의 관계, 계층을 분석하여 귀납적인 범주를 만드는 방법이다.[23] 본 연구에서는 도덕적 추론 과정으로부터 추상적인 개념들을 추출하고 체계화하기 위하여 코딩을 통한 프로토콜 분석을 수행한다.

4. 시나리오 설계

4-1. 교통사고 데이터베이스 사례 조사

실험을 위한 교통사고 시나리오는 미국 도로교통안전국(The National Highway Traffic and Safety Administration; NHTSA)이 게시하는 특별 교통사고 조사(Special Crash Investigation; SCI) 데이터베이스의 자료를 바탕으로 구성되었다.[24] SCI 데이터베이스는 약 1,500 건에 달하는 전문적 교통사고 분석을 제공한다. 각 사례는 교통사고의 배경(도로 환경, 교통 정보, 차량 및 탑승자), 과정(순차적 충돌, 운전자의 행동), 결과(탑승자의 부상 정보 등)를 상세히 다룬다.

실험 시나리오는 도덕적 가치의 충돌로 딜레마가 일어나고, 자율주행 자동차가 피할 수 없는 교통사고를 배경으로 설계되어야 한다. 따라서 저자들은 논의를 통하여 시나리오의 참고사례 선정 조건을 교통사고 배경, 과정, 결과의 측면에서 설정하였다. 배경에 있어서는 주변에 장애물이 많아 충돌을 피할 공간이 없었던 경우이다. 과정에 있어서는 다른 차량의 돌발적 행동이 사고의 원인이 된 경우이다. 결과에 있어서는 사망할 수도 있는 심각한 부상 피해가 일어난 경우이다. SCI 데이터베이스에서 세 조건에 맞는 사례를 검색하였다.

조건에 맞는 사례들을 비슷한 유형끼리 분류한 결과, 교차로에서 일어난 측면충돌, 중앙선을 침범한 차량에 의한 정면충돌, 교통신호에 따른 정차 중 후방 차량에 의한 후면충돌의 유형으로 나타났다. 이 때, 자율주행 자동차는 정차 중 충돌을 예측하고 회피할 수 있기 때문에 후방충돌 유형을 제외한 측면충돌, 정면충돌 유형의 사례들을 시나리오의 참고사례로 설정하였다.

4-2. 교통사고 속 도덕적 딜레마 시나리오 설계

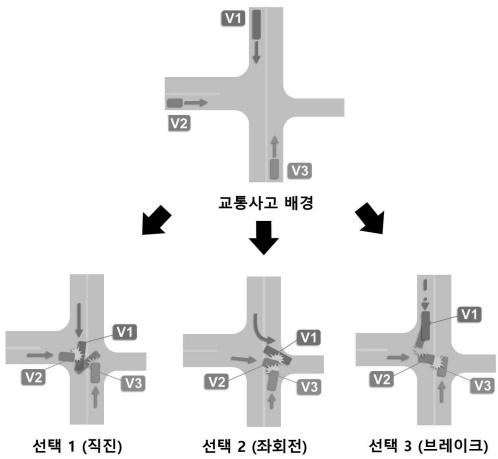
SCI 데이터베이스에서 선정한 유형의 사례들을 주제로 파일럿 인터뷰를 수행하였다.[25][26][27] 파일럿 인터뷰는 운전자 입장의 도덕적 갈등 요인과 실험의 타당성 확인을 목적으로, 여섯 명의 운전 경험자들과의 심층 인터뷰로 진행되었다. 그 결과, 교통사고 시 공리주의적 원칙뿐만 아닌 다양한 가치요인이 존재한다는 것을 확인하였다 (탑승자에 대한 책임감, 교통신호에 대한 의무, 보행자에 대한 보호의식 등). 또 참여자들에게 공통적으로 보행자에 대한 보호의식을 관찰할 수 있었다. 이에 보행자 시나리오의 필요성이 확인되었다.

파일럿 인터뷰를 통해 확인한 가치요인과 통찰을 바탕으로 총 네 가지 실험 시나리오를 설계하였다. 시나리오는 교통사고의 배경, 과정, 결과로 구성하였다. 사고 배경은 인터뷰 자료로 사용한 SCI 교통사고 사례를

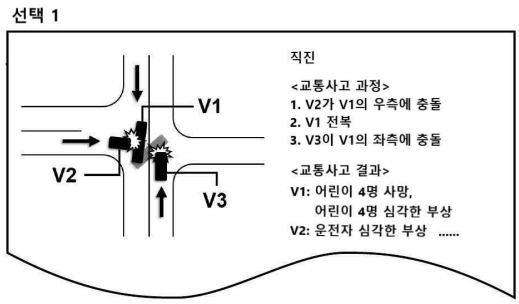
20) Fonteyn, M. E., Kuipers, B., & Grobe, S. J., A description of think aloud method and protocol analysis, Qualitative health research, Vol.3, No.4, 1993, p.430-441.
21) Ericsson, K. A., & Simon, H. A., Verbal reports as data, Psychological review, Vol.87, No.3, 1980, p.215.
22) Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N., The psychology of moral reasoning, Judgment and Decision making, Vol.3, No.2, 2008, p.121.
23) Bohm, A., A companion to qualitative research, Sage Publications, 2004.
24) NHTSA (National Highway Traffic and Safety Administration), Special Crash Investigations, https://www.nhtsa.gov/research-data/special-crash-investigations-sci, n.d.
25) Indiana University Transportation Research Center, On-Site School Bus Investigation (Case No. IN08002), 2009.
26) Crash Research & Analysis, Inc., On-Site Ambulance Crash Investigation (Case No. CR120002), 2013.
27) Calspan Crash Data Research Center, Calspan Remote Certified Advanced 208-Compliant Vehicle Crash Investigation (Case No. 2004-02-041A), 2006.

기반으로 하였다. 각 시나리오의 사고 과정은 두, 세 가지 선택지로 나뉘는데, 선택지마다 다른 피해 결과가 주어진다 [그림 1]. 피해 정보는 사고기준 단일화 모델(Model Minimum Uniform Crash Criteria)의 부상 등급을 참고하여, 피해자마다 사망, 심각한 부상, 심각하지 않은 부상, 경미한 부상의 등급으로 표기하였다. 각 선택지는 파일럿 인터뷰에서 관찰한 운전자 가치요인의 충돌을 반영하여 도덕적 갈등을 일으킬 수 있는 과정과 결과로 설정하였다. 횡단보도가 있는 시나리오에서는 보행자가 등장한다는 가정을 추가하였다.

예를 들어, 시나리오 <S1>의 배경은 27명의 어린이가 탑승한 통학버스(V1) 운전자가 교차로를 지나려 할 때, 우측에서 달려오는 소형차(V2)로 인해 사고 직전에 처한 상황이다 [그림 1]. 시나리오를 바라보는 추론자(reasoner)는 V1의 운전자 입장이며, V1이 제안하는 행동들과 결과 예측을 보고 한 가지를 선택한다. [그림 2]는 선택에 따른 사고 과정 및 결과를 보여준다. V1이 직진하면 V2와 충돌하고, 8명의 어린이가 사망 및 중상을 입는다. V1이 좌회전 하면, V2가 V1 및 지나가던 차량(V3)과 충돌한 결과, 3명의 어린이가 심각한 부상을 입고 V2의 운전자가 사망한다. V1이 브레이크를 밟으면, V2와 V3이 충돌의 중심이 되어 V3의 운전자가 사망하고 2명의 어린이가 심각한 부상을 입는다. 이 시나리오는 좌회전과 브레이크 사이에서 어린 승객의 심각한 부상과 무고한 타인(V3 운전자)의 죽음에 대한 갈등을 유도한다.



[그림 1] 시나리오의 교통사고 배경 및 과정 선택지



[그림 2] 시나리오 선택지의 교통사고 과정 및 결과

본 글에서는 지면의 한계로 <S1>을 예시로 시나리오의 핵심 구성을 설명하였다. 실험에서 사용한 네 가지 시나리오는 http://hdl.handle.net10203/266047에서 열람할 수 있다.[28)]

5. 실험 및 분석

5-1. 시나리오에 대한 도덕적 추론 실험

본 실험은 사고 실험으로, 참여자가 교통사고의 도덕적 딜레마 시나리오를 보며 의사결정을 내리는 인지적 과정을 수집하였다. 그 도구로서 사고 구술과 자유응답 인터뷰(open-ended interview)를 사용하였다.

참여자는 소셜 네트워크 서비스와 지역 커뮤니티에서 공개 방식과 비공개 방식으로 모집되었다. 참여자들에게는 1만원의 보상이 지급되었다. 성인 이상의 다양한 연령대, 운전 경험, 직업의 참여자 33명이 표본으로 구성되었으며, 성별은 남자 12명, 여자 21명이었다.

실험환경은 연구실과 온라인 환경으로 병행하여 구성하였다. 연구실에서는 시나리오 다이어그램 카드를 보며 면대면 실험을 진행하였다. 온라인에서는 슬라이드로 다이어그램을 전달하고 스카이프로 진행하였다.

실험이 시작되면 시나리오의 다이어그램을 시각 자료로 제공하고, 참여자가 시나리오의 배경과 과정, 결과를 이해하는 과정부터 구술을 시작하도록 하였다. 시나리오에 대한 설명과 질의로 참여자의 이해를 도왔다. 참여자가 추론 활동을 수행하여 의사결정을 마치면, 시나리오의 변수를 하나씩 바꾸어 개인적 가치요인에 대

28) 이기쁨, 자율주행 자동차의 도덕적 의사결정을 위한 인간의 도덕적 추론 분석, 한국과학기술원 석사학위논문, 2018, p.29-33.

한 상세한 구술을 유도하였다. 참여자의 구술에서 가치 요인에 대한 정보가 부족할 경우 가치 중립적 입장에서 참여자의 의견을 이끌어내는 질문을 던졌다. 이와 같은 절차를 네 가지 시나리오에 대하여 수행하여, 평균적으로 45분의 시간이 소요되었다. 모든 과정을 녹음하여 구술 자료를 수집하였다. 참여자의 몸짓, 표정, 눈길과 같은 비언어적 정보는 현장에서 필기하였다.

5-2. 프로토콜 분석을 통한 도덕 코드 추출

음성과 메모로 기록된 자료를 프로토콜로 변환하였다. 참여자 별로 사고 구술 및 인터뷰한 음성 자료를 문자로 전사하였다. 이때 음성 자료를 반복적으로 들으며 참여자의 말투와 의도를 포착하였다. 이는 참여자의 비언어적 표현과 함께 참여자의 감정에 대한 자료로서 문장 끝에 삽입하였다 [표 1-(a)].

다음으로, 프로토콜에 대한 질적 분석으로 도덕적 가치요인을 코딩하였다. [표 1]은 참여자(P1)의 〈S1〉 프로토콜에 대한 코딩 결과의 일부분이다. 먼저 프로토콜에서 강조되거나 반복적으로 언급된 키워드를 찾았다. 키워드를 통해 추상적인 코드를 임시로 명명하고, 이들을 비교하여 비슷한 코드를 통합하고 복합적인 코드를 분리하는 과정을 반복하였다. 코드의 이름은 그 개념을 표현할 수 있도록 계속하여 수정되었다.

[표 1] 프로토콜 코딩의 결과물

선택지	사고 구술	코딩 결과
브레이크 / 좌회전	"내가 태운 어린이 한 명의 심각한 부상과 사람 생명을 저울질하고 있어요."	동승자 보호, 중상, 일반차량 보호
	"브레이크 밟았을 때와 좌회전했을 때의 과실에 대해서도 생각하고 있어요."	패널티
브레이크	"브레이크를 밟는 게 맞는 것 같기도 하고."	
	"그건 왜죠? 과실 측면에서인가요?"*	
	"네, 가장 방어운전인 것 같은데… V3." (멈추고 고민)[a]	방어운전

*연구자의 질문

코드의 신뢰도를 위하여 두 명의 저자가 개별적으로 코딩과 코드 개선을 진행한 뒤, 코드의 평정자 간 신뢰도(Inter-rater reliability)를 평가하였다. 그 결과, 평정자 간 코드가 코헨의 카파 상관계수 K=0.35 수준으로 일치하였다. 이후 친화도 다이어그램(affinity diagram) 분석으로 코드를 클러스터링하여 계층을 분류하였다.

6. 결과

운전자의 도덕 코드는 규범적 요인(normative ideas), 절차적 요인(procedural ideas), 행동적 요인(actional ideas)에 따라 분류되었다. 본 장은 요인별로 코드의 정의와 빈도를 보여준다. 코드의 빈도는 해당 코드를 고려하여 의사결정을 내린 참여자의 수를 의미한다. 참여자의 프로토콜 데이터에서 해당 코드가 추출되었는지를 기준으로 개수하였다.

6-1. 운전자의 규범적 도덕 코드

규범적 도덕 코드는 사회적으로 올바르게 받아들여지는 행동양식과 그러한 행동으로 유도하는 가치요소를 가리킨다. 즉, 사람들이 해야 할 것과 하지 말아야 할 것으로 여기는 코드들이 해당한다. 이러한 코드들은 책임, 행위, 정책이라는 요소와 관련이 있었다 [표 2]. 책임에 관한 코드는 역할과 행동에 따르는 당위성에 의한 요인들이다. 행위에 관한 코드는 하지 말아야 할 행위를 판단할 수 있는 속성들이다. 정책에 관한 코드는 사회적 합의를 이루었거나 지향하는 규범들이다.

규범적 도덕 코드에서는 사상자 최소화(91%), 과실 책임(79%), 교통법규 준수(70%)가 가장 높은 빈도를 보였다. 다음은 빈도가 높은 코드의 인용문이다.

사상자 최소화: "일단 사망자가 없는 선택지는 없고 누군가 죽어야 한다면 한 명이 되어야 하지 않을까 싶어요." (P17)

과실 책임: *"V2는 빨간 불이 있는데 직진을 한 거잖아요? 굳이 따지자면 책임은 V2에 있으니까, 그 둘 중에, V2한테 책임이 좀 더 전가되는 V2 사망 쪽으로 선택을 하게 될 것 같아요."* (P10)

교통법규 준수: *"그렇다고 내가 좌회전을 하자니 내 신호도 아닐뿐더러."* (P9)

[표 2] 규범적 도덕 코드의 의미와 빈도

코드 (빈도)	의미
책임	
과실 책임 (26/33[79%])	사고의 원인을 제공한 사람에게 피해와 태만에 대한 책임이 있다.
운전자 책임 (19/33[58%])	운전하는 차량을 사고로부터 보호하여 피해를 최소화하는 것이 운전자의 책임이다.
특수차량 책임 (14/33[42%])	특정 목적을 가지고 운행되는 차량은 역할에 부합하는 행동을 해야 하며, 그 운전자에게는 차량의 행동에 대한 책임이 있다.

행위	
직접적 가해 (18/33[55%])	다른 사람에게 물리적인 접촉을 통해 피해를 입히는 직접적 행위를 지양한다.
의도적 가해 (13/33[39%])	다른 사람이 입을 피해를 알면서 피해를 입히는 의도적 행위를 지양한다.
적극적 가해 (14/33[42%])	행동을 취함으로써 다른 사람에게 피해를 입히는 적극적 행위를 지양한다.
정책	
사상자 최소화 (30/33[91%])	전체적 인명피해를 최소화해야 한다.
교통법규 준수 (23/33[70%])	신호, 속도제한, 추월, 보행자 통행 등의 도로교통법을 의식하고 지켜야 한다.
방어운전 (12/33[36%])	운전하는 차량의 안전상태를 유지하기 위하여 조심스러운 운전을 해야 한다.

6-2. 운전자의 절차적 도덕 코드

절차적 도덕 코드는 도덕적 추론 중간에 작용하는 주관적 관점이나 감정을 가리키는 것으로, 규범 요인과 행동 요인의 다리 역할을 한다. 같은 규범 요인을 가지더라도 주관적인 절차 요인의 작용에 따라 다른 행동 요인에 무게를 실을 수 있다. 절차 요인은 감정, 불확실성, 피해에 관한 코드로 나누어진다 [표 3]. 감정에 관한 코드는 추론 중 나타나는 도덕적 감정을 가리킨다. 불확실성에 관한 코드는 불확실한 사고의 결과에 대한 태도이다. 피해에 관한 코드는 사람의 피해 정도를 파악하는 주관적 방식이다.

절차적 도덕 코드 중에서는 중상(88%), 죄책감(76%), 낙관적 예측(70%)이 가장 높은 빈도를 보였다. 다음은 이 코드들의 인용문이다.

중상: *"평생 오는 장애로 인한 피해도 다 내가 떠안아야 하는 거잖아요. 이건 또 다치고 안 다치고의 문제가 아니라 삶의 고통이 따라오는 거네요."* (P3)

죄책감: *"내 중심적일 수도 있는데, 내가 사람을 쳐서 죽이고 싶진 않아요. (...) 얼마나 힘들겠어요. 삶을 살아갈 수 있을까 싶을 정도로."* (P15)

낙관적 예측: *"두 사람이 다치지만 그래도 살 확률도 50%니까, 이걸 봤죠. 결과적으로."* (P2)

[표 3] 절차적 도덕 코드의 의미와 빈도

코드 (빈도)	의미
감정	
죄책감 (25/33[76%])	자신의 행동과 그 결과에 대한 미안함과 책임감을 느낀다.
공감 (22/33[67%])	피해자의 입장과 고통을 자신의 것처럼 생각하고 연민을 느낀다.
분노 (18/33[55%])	불합리한 피해자가 발생하는 상황에 대해 부정적 감정을 느낀다.
불확실성	
낙관적 예측 (23/33[70%])	예측된 수치보다 나은 결과와 선택의 이점(benefit)을 기대한다.
비관적 예측 (13/33[39%])	예측된 수치보다 나쁜 결과와 선택의 위험성(risk)을 고려한다.
통제성 (9/33[27%])	불확실한 미래 중 자신이 통제할 수 있는 영역에 집중한다.
피해	
중상 (29/33[88%])	심각한 부상은 사망이나 장애 등의 사후 고통을 수반할 가능성이 있기 때문에 사망과 견줄 수 있는 큰 피해이다
부상 (21/33[64%])	심각하지 않은 부상은 사망과 비교하여 작고 수용 가능한 피해이다.
피해 수치화 (20/33[61%])	인적 피해의 규모를 수학적으로 평가한다.
패널티 (9/33[27%])	사고 이후에 따르는 정신적, 법적, 물질적 피해도 사고로 인한 피해이다.

6-3. 운전자의 행동적 도덕 코드

행동적 도덕 코드는 어떤 대상을 보호하겠다는 목적으로서 구체적인 행동의 방아쇠가 된다. 행동 요인은 개개인에게 내재되어 있으며, 문제 상황과 규범, 절차 요인에 의해 무게가 조절되기도 한다. 코드들은 방어, 이타성, 보살핌과 관련이 있다 [표 4]. 방어에 관한 코드는 추론자 자신과 같은 차량공간의 대상에 관한 방어적 보호 의식으로, 운전차량 보호라고 표현할 수 있다. 이타성에 관한 코드는 추론자 자신과 가까운 공동체가 아닌 대상을 보호하려는 태도이다. 보살핌에 관한 코드는 사고 상황에서 상대적으로 취약한 조건을 가진 대상을 보호하려는 태도이다.

행동적 도덕 코드 중 가장 빈도수가 높은 코드는 동승자 보호(94%), 자기 보호(88%), 보행자 보호(79%)였다. 친지 보호는 아닌 연구자의 선택적 질문으로 얻은 코드이다. 따라서 친지가 사고에 관련되는 시나리오로 실험을 진행한다면 더 높은 빈도를 얻을 것으로 예상된다. 다음은 빈도가 높은 코드들의 인용문이다.

동승자 보호: *"제 차에 타고있는 사람이 우선이라고 생각해요. 어린이가 타든 노인이 타든 제가 제 차에 타고있는 사람을 먼저 구해야 한다고 생각해요."* (P30)

자기 보호: *"나라면 일단 내가 살아야 되는 방향으로 생각을 하겠지."* (P25)

보행자 보호: *"보행자는 자기를 보호해 줄 수 있는 게 없는데 차로 치는 게 좀 위협적인 것 같아서."* (P4)

[표 4] 행동적 도덕 코드의 의미와 빈도

코드 (빈도)	의미
방어 [운전차량 보호]	
동승자 보호 (31/33[94%])	같은 차량에 탑승하고 있는 사람은 보호 대상이다.
자기 보호 (29/33[88%])	운전자 자신은 보호받아야 하는 대상이다.
친지 보호 (14/33[42%])	가족, 친구, 동료 등 사회적 관계가 있는 사람은 보호 대상이다.
이타성	
일반차량 보호 (23/33[70%])	사고에 관여된 차량은 모두 보호 대상이다.
과실차량 보호 (22/33[67%])	과실을 범한 차량도 보호 대상에 포함된다.
승합차량 보호 (6/33[18%])	많은 사람이 탑승할 가능성이 있는 승합차류는 보호 대상이다.
보살핌	
보행자 보호 (26/33[79%])	보행자는 보호 대상이다.
어린이 보호 (24/33[73%])	어린이는 보호 대상이다.

7. 결론

7-1. 결론 및 논의

본 연구는 실제 교통사고 사례에 기반한 도덕적 딜레마 시나리오를 고안하고, 시나리오의 운전자 시점에서 사고 구술 실험을 진행하여 사람의 도덕적 의사결정과정을 관찰하였다. 그 결과, 운전자의 입장에서 교통사고 속 도덕적 의사결정의 근거가 되는 도덕 코드를 추출하였다.

32개의 도덕 코드는 자율주행 자동차의 구체적 딜레마에 연관된 도덕적 요인의 다양성을 보여준다. 제삼자가 아닌 도덕 행위자로서 사람들은 공리주의뿐 아니라 운전자의 책임, 교통법규 준수, 죄책감 등 다양한 요인을 토대로 추론하였다. 집단이 공유하는 도덕 코드는 사회적, 문화적 차이에 따라 다르기 때문에 일반화에는 주의가 필요하다. 그럼에도 본 연구가 제시한 도덕 코드들의 군집(책임, 행위, 정책, 감정, 불확실성, 피해, 방어, 이타성, 보살핌)은 사람의 도덕적 추론에 있어 잘 알려진 요인들과 일치한다.[29][30] 또한 도덕 코드의 최상위 군집(규범 요인, 절차 요인, 행동 요인)은 절차적 측면과 규범 요소를 고려하여 의사결정을 내리는 디자인 씽킹(design thinking) 방법론과 맥락을 같이 한다.[31] 세 군집의 이름은 이를 토대로 명명한 것이다.

도덕 코드의 다양한 측면은 도덕적 의사결정에 대한 프로토콜 분석을 통해 다시 한번 통찰을 얻을 수 있었다. 첫째, 두 사람이 같은 사고 결과를 선택하더라도, 그들 각자가 왜, 어떻게 그와 같은 선택을 내렸는지는 서로 다른 가치요인의 작용이었다. 둘째, 두 사람이 같은 규범에 동의하더라도, 다양한 절차, 행동 요인들에 의하여 반대의 의사결정을 내릴 수도 있다. 따라서, 도덕적 행위자의 역할을 수행할 자율주행 자동차는 단순히 특정 윤리원칙을 따르는 것이 아니라, 현실 세계에서 실제로 사용되는 가치요인들을 반영한 의사결정을 내려야 할 것이다.

학술적 측면에서 본 연구가 제안하는 코드와 계층구조는 자율주행 자동차의 윤리 연구에 인간의 도덕성 관점에서 질적 개념들을 제시하였다. 본 연구의 결과는 선 연구된 대중의 의사결정 데이터를 보충하는 "의사결정의 이유"에 대한 기초자료로서 의미가 있다. 본 연구는 모럴 머신과 달리 도덕 행위자인 운전자 입장에서의 도덕적 추론에 초점을 맞추었는데, 이러한 방식은 행위자의 추론을 관찰하여 도덕성을 모델링하는 논리 프로그래밍(logic programming)의 패러다임과도 맥을 같이 한다.[32]

산업적 측면에서 본 연구는 자율주행 자동차의 도덕 시스템 설계자에게 구체적인 설계 요소를 제공한다는 의의가 있다. 운전자의 도덕 코드는 자율주행 자동차의 도덕적 의사결정을 계획하고 논의하기 위한 구체적인 용어로 활용될 수 있다. 또한 코드의 계층구조를 이용하여 규범, 절차, 행동의 측면에서 체계적인 의사결정 과정을 설계할 수 있다. 이러한 설계가 사람의 입장에서 풀어낸 코드를 기반으로 하기에, 자율주행 자동차의 의사결정 근거에 대한 사람의 긍정적 공감을 촉진할 것으로 기대할 수 있다.[33]

29) Wildermuth, C. et al., Op. cit., p.17-42.
30) Cushman, F. et al., Op. cit., p.1082-1089.
31) Rowe, P. G., Design Thinking. MIT press, 1991.
32) Saptawijaya, A., & Pereira, L. M., Towards Modeling Morality Computationally with Logic Programming, International Symposium on Practical Aspects of Declarative Languages, 2014, pp.104-119.
33) Rhim, J., Social-Value Embedded Ethical Decision-Making Framework of Autonomous Vehicles: A Cross-Cultural Study, Korean Advanced Institute of Science and Technology Ph.D. dissertation, 2020, p.29-30.

7-2. 향후 연구 방향

본 연구가 제시하는 도덕 코드들은 교통사고 속 도덕적 의사결정의 이유가 되는 가치요인들을 명명한 의미가 있다. 그럼에도 불구하고, 도덕 코드들을 투명한 도덕적 의사결정 시스템에 활용하려면 그들 간의 상관관계에 대한 연구가 진행되어야 한다. 따라서 도덕 코드 간의 관계를 가시화하고, 인공지능 시스템이 학습 데이터로 사용할 수 있도록 자료화하는 연구를 제안한다. 또한 후속 연구에서는 재현이 가능한 체계적 코딩 절차를 개발하고, 감정 분석 도구를 사용함으로써 도덕적 추론에 대한 객관적 코딩을 보완할 수 있다.

참고문헌

1. 이기쁨, 자율주행 자동차의 도덕적 의사결정을 위한 인간의 도덕적 추론 분석, 한국과학기술원 석사학위논문, 2018.
2. Bohm, A., A companion to qualitative research, Sage Publications, 2004.
3. Rowe, P. G., Design Thinking. MIT press, 1991.
4. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I., The moral machine experiment, Nature, Vol.563, No.7729, 2018.
5. Bonnefon, J. F., Shariff, A., & Rahwan, I., The social dilemma of autonomous vehicles, Science, Vol.352, No.6293, 2016.
6. Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N., The psychology of moral reasoning, Judgment and Decision making, Vol.3, No.2, 2008.
7. Cushman, F., Young, L., & Hauser, M., The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm, Psychological science, Vol.17, No.12, 2006.
8. Danielson, P., Surprising judgments about robot drivers: Experiments on rising expectations and blaming humans, Etikk i praksis-Nordic Journal of Applied Ethics, Vol.9, No.1, 2015.
9. Dilich, M. A., Kopernik, D., & Goebelbecker, J., Evaluating driver response to a sudden emergency: Issues of expectancy, emotional arousal and uncertainty, SAE Transactions, 2002.
10. Ericsson, K. A., & Simon, H. A., Verbal reports as data, Psychological review, Vol.87, No.3, 1980.
11. Fonteyn, M. E., Kuipers, B., & Grobe, S. J., A description of think aloud method and protocol analysis, Qualitative health research, Vol.3, No.4, 1993.
12. Goodall, N. J., Ethical decision making during automated vehicle crashes, Transportation Research Record, Vol.2424, No.1, 2013.
13. Greene, J. D., Our Driverless Dilemma, Science, Vol.352, No.6293, 2016.
14. Itoh, M., Lemoine, M., Robache, F., & Morvan, H., An analysis of driver's avoiding maneuver in a highly emergency situation, SICE Journal of Control, Measurement, and System Integration, Vol.8, No.1, 2015.
15. Lin, P., Why Ethics Matters for Autonomous Cars, Autonomous Driving, 2016.
16. Saptawijaya, A., & Pereira, L. M., Towards Modeling Morality Computationally with Logic Programming, International Symposium on Practical Aspects of Declarative Languages, 2014.
17. Waldrop, M., A World of Driverless Cars, Nature, Vol.518, No.7537, 2015.
18. Wildermuth, C., e Souza, C. A. D. M., & Kozitza, T., Circles of ethics: the impact of proximity on moral reasoning, Journal of business ethics, Vol.140, No.1, 2017.
19. Rhim, J., Social-Value Embedded Ethical Decision-Making Framework of Autonomous

Vehicles: A Cross-Cultural Study, Korean Advanced Institute of Science and Technology Ph.D. dissertation, 2020.

20. Calspan Crash Data Research Center, Calspan Remote Certified Advanced 208-Compliant Vehicle Crash Investigation (Case No. 2004-02-041A), 2006.

21. Crash Research & Analysis, Inc., On-Site Ambulance Crash Investigation (Case No. CR120002), 2013.

22. Indiana University Transportation Research Center, On-Site School Bus Investigation (Case No. IN08002), 2009.

23.https://www.nhtsa.gov/research-data/special-crash-investigations-sci.