

TRUSTWORTHY AI

인공지능과 어떻게 공존할 것인가

클라우드나인
CLOUDNINE

TECH
FRONTIER

인공지능의 신뢰성 - 데이터를 중심으로

2021. 10

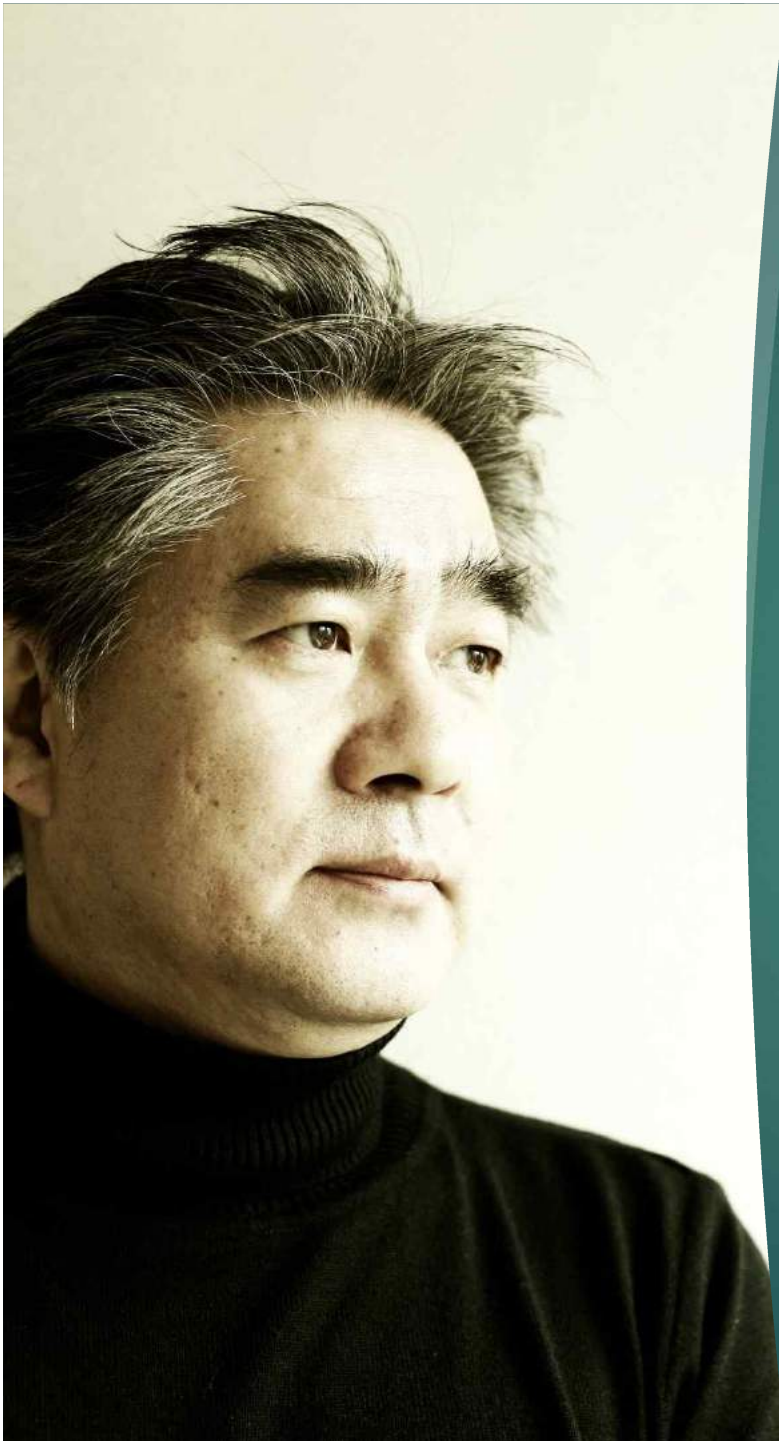
한상기 /

테크프론티어 대표

발표자 소개

2

- ▶ "서울대 컴퓨터공학과를 졸업하고, 카이스트에서 인공지능 분야 중 지식 표현에 관한 연구로 박사 학위를 취득했다. 삼성전자 전략기획실과 미디어 서비스 사업팀에서 인터넷사업을 담당한 후, 2003년 다음커뮤니케이션 전략대표와 일본 법인장을 역임했다. 두 번의 창업을 했으며, 카이스트와 세종대학교 교수를 거쳐 2011년부터 테크프론티어 대표를 맡고 있다.
- ▶ 현재 기업을 위한 기술 전략컨설팅, 정부 정책 자문과 연구 수행 그리고 기술과 사회에 관한 강연을 하고 있다. 공공 영역에서는 인공지능 데이터셋 구축을 위한 AI 데이터 로드맵 총괄기획위원 등의 활동을 하고 있다.
- ▶ AI 타임즈, NIA AI Data Insight, KISA REPORT, NIA Digital Service 이슈 리포트 등 여러 매체에 기술 관련 칼럼을 기고하고 있다.
- ▶ 저서로는 '한상기의 소셜미디어 특강', '인공지능은 어떻게 산업의 미래를 바꾸는가', '4차 산업혁명과 미래사회', '초연결시대 인간-미디어-문화', '2019 미래를 읽다' 등 다양한 공저가 있다."



인공지능의 신뢰성에 대한 계속되는 문제 발생

3

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.



Google's algorithm for detecting hate speech is racially biased



AI systems meant to spot abusive online content are far more likely to label tweets "offensive" if they were posted by people who identify as African-American

Police across the US are training crime-predicting AIs on falsified data

A new report shows how supposedly objective systems can perpetuate corrupt policing practices.

by Karen Hao

February 13, 2019

In May of 2010, prompted by a series of high-profile scandals, the m Orleans asked the US Department of Justice to investigate the cit

TECHNOLOGY NEWS · OCTOBER 11, 2018 / 8:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows a physical perturbation applied to a Stop sign. The systems classify the sign on the right as a Speed Limit: 45 mph sign! Source: Robot: Physical-World Attacks on Deep Learning Visual Classification.

Dutch court rules AI benefits fraud detection system violates EU human rights

SyRI was used to predict who may be at high risk of conducting housing or social security fraud.

NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

Heidi Ledford

TECH
FRONTIER



안전의
문제도
점점
심각해지고
있음

AI의 신뢰, 책임, 윤리의 문제는 계속 발전되어 왔음

- ▶ Friendly AI (Eliezer Yudkowsky, 2001)
- ▶ Beneficial AI (FLI, 2014)
- ▶ Trustworthy AI (EC, 2019)
- ▶ Responsible AI – Google, Microsoft, Facebook
- ▶ 80개 이상의 원칙과 가이드라인 – EC, OECD, UN, 바티칸, 각 나라별 원칙, 기업의 원칙 등

PRINCIPLED ARTIFICIAL INTELLIGENCE

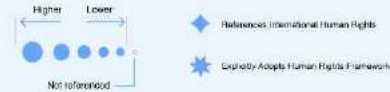
A Map of Ethical and Rights-Based Approaches to Principles for AI

Authors: Jessica Fjeld, Nele Achten, Hannah Hillgoss, Adam Nagy, Madhulika Srikumar
 Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

HOW TO READ:

Date, Location
Document Title
 Actor

COVERAGE OF THEMES:



The size of each dot represents the percentage of principles in that theme contained in the document. Since the number of principles per theme varies, it's informative to compare dot sizes within a theme but not between themes.

The principles within each theme are:

Privacy:

- Privacy
- Control over Use of Data
- Consent
- Privacy by Design
- Recommendation for Data Protection Laws
- Ability to Rectify Processing
- Right to Rectification
- Right to Erasure

Accountability:

- Accountability
- Recommendation for New Regulations
- Impact Assessment
- Evaluation and Auditing Requirement
- Verifiability and Replicability
- Liability and Legal Responsibility
- Ability to Appeal
- Environmental Responsibility
- Crises of a Missing Body
- Remedy for Automated Decision

Safety and Security:

- Safety
- Safety and Reliability
- Predictability
- Security by Design

Transparency and Explainability:

- Explainability
- Transparency
- Open Source Data and Algorithms
- Notification when Interacting with an AI
- Notification when AI Makes a Decision about an Individual
- Regular Reporting Requirement
- Right to Information
- Open Procurement (for Government)

Fairness and Non-discrimination:

- Non-discrimination and the Prevention of Bias
- Fairness
- Inclusiveness in Design
- Inclusiveness in Impact
- Representative and High Quality Data
- Equality

Human Control of Technology:

- Human Control of Technology
- Human Review of Automated Decision
- Ability to Opt out of Automated Decision

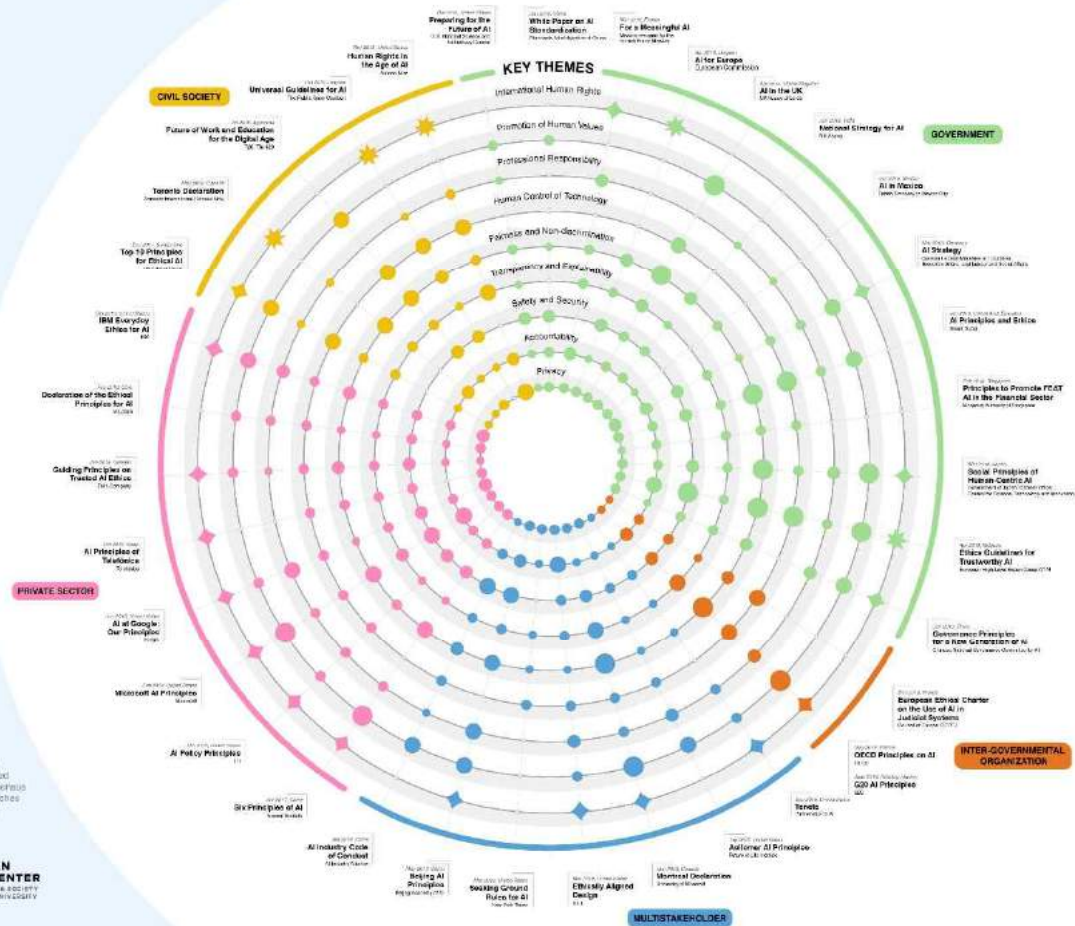
Professional Responsibility:

- Responsible Design
- Consideration of Long-Term Effects
- Accuracy
- Scientific Integrity

Promotions of Human Values:

- Leveraged to Benefit Society
- Human Values and Human Flourishing
- Access to Technology

Further information on findings and methodology is available in Principled Artificial Intelligence: Mapping Convergence in Ethical and Rights-Based Approaches (Berkman Klein, 2023) available at berkman.klein.org.

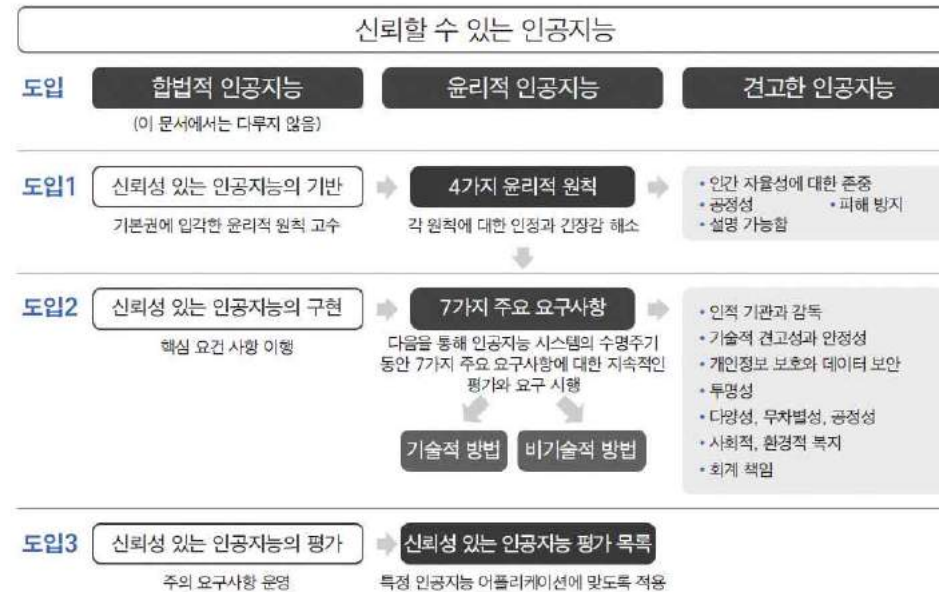


MULTISTAKEHOLDER

신뢰할 수 있는 인공지능: 새로운 거버넌스 논의 (2019)

- ◆ ETHICS GUIDELINES FOR TRUSTWORTHY AI (EC High-Level Expert Group on AI)

유럽연합 집행위원회가 제시한 '신뢰할 수 있는 인공지능을 위한 프레임워크'



EU의 AI Law 초안 (2021. 4)

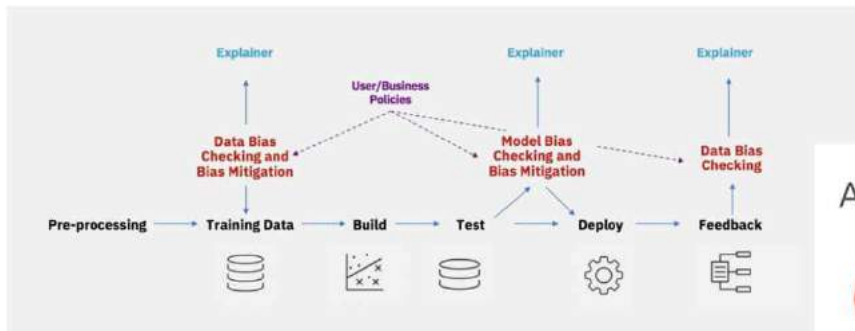
8

- ▶ 용납할 수 없는 위험, 고위험, 제한된 위험, 최소 위험이라는 네 범주로 좀 더 명확히 분류
- ▶ 용납할 수 없는 위험에는 사람들의 안전, 생활, 권리에 명확한 위험이 되는 인공지능 시스템, 사람들의 자유 의지를 방해하면서 행동을 조작하는 시스템
 - 미성년자의 위험한 행동을 유발하는 음성 비서를 이용하는 장난감이나 정부가 사람들을 평가하는 '사회적 점수 매기기' 같은 시스템
- ▶ 고위험
 - 시민의 생명과 건강을 위험하게 만들 수 있는 중요 인프라 (예: 교통)
 - 학습이나 전문 과정에 대한 접근을 결정하는 교육이나 직업훈련
 - 제품의 안전 관련 부품 (예: 로봇 수술에서의 인공지능 애플리케이션)
 - 고용, 노동자 관리, 자영업에 대한 접근 (예: 채용 과정을 위한 이력서 분류 소프트웨어)
 - 필수적인 민간 또는 공공서비스 (예: 대출받을 기회를 부정하는 신용 점수 매기기)
 - 시민의 기본권을 침해할 수 있는 법 집행 (예: 증거 신뢰도에 대한 평가)
 - 이민, 망명, 국경 통제 관리 (예: 여행 문서의 진위 검증)
 - 사법 행정과 민주적 절차 (예: 구체적인 사실에 대한 법 적용)
- ▶ 제한된 위험에 속하는 인공지능 시스템은 특정한 투명성 의무를 진다.
 - ▶ 법을 위반했을 때에는 3천만 유로 또는 글로벌 매출의 6%까지 (둘 중에 더 큰 쪽을 적용함) 벌금

공정성 (Fairness)

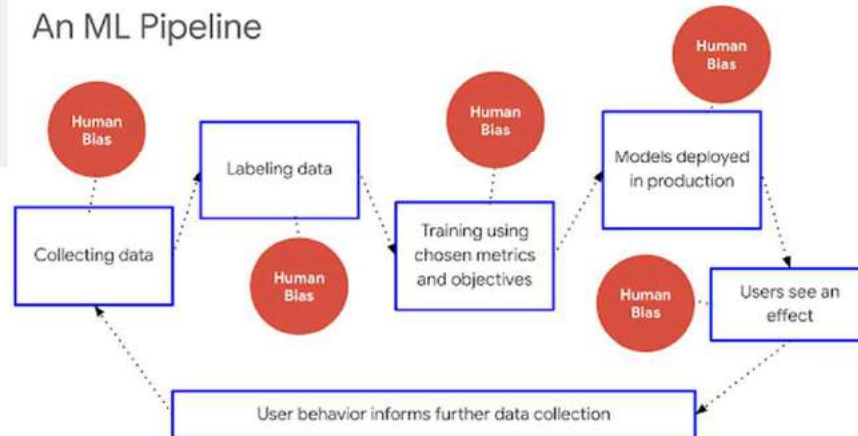
공정성 문제는 AI/ML의 모든 과정에서 발생할 수 있다

10



September 19, 2018 | Kush R. Varshney (IBM)

Google, December 11, 2019



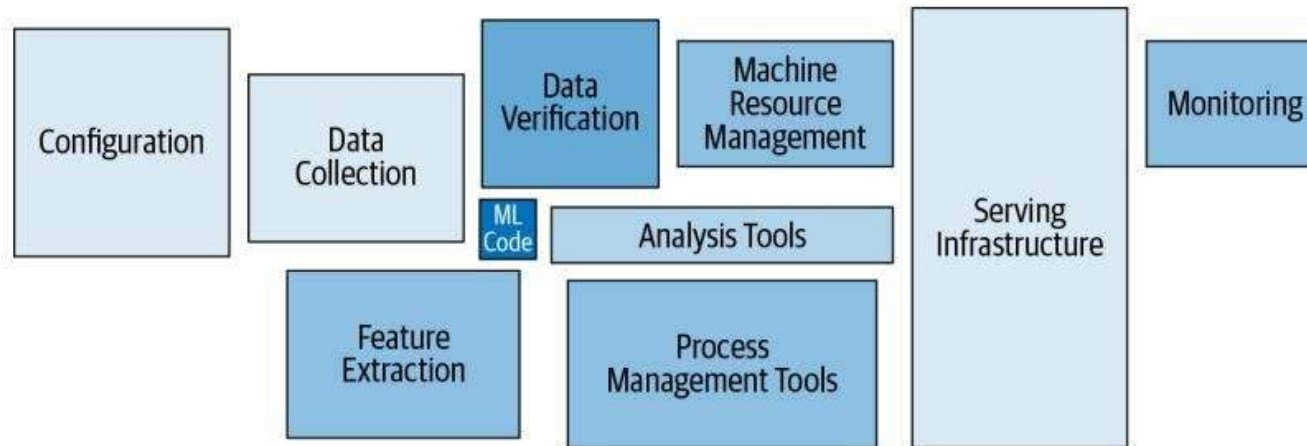
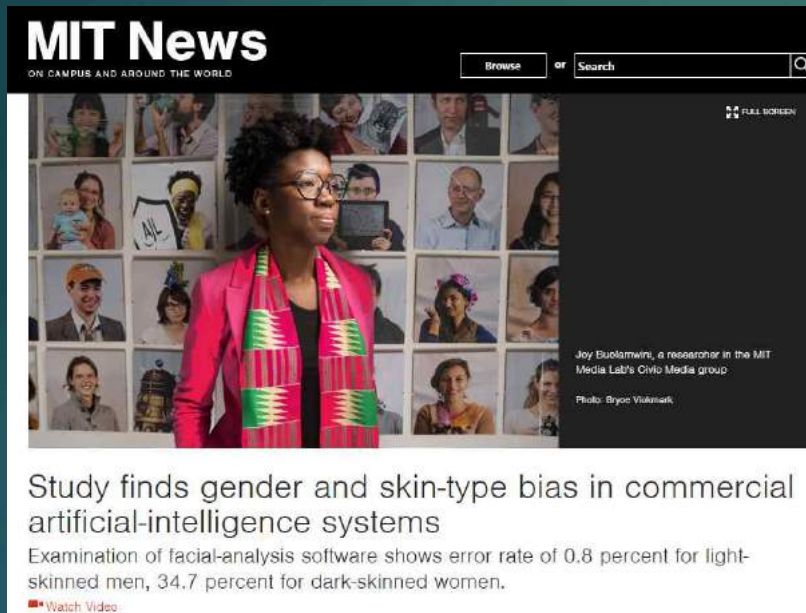


Figure 8-2. Code is only a fraction of real-world machine learning systems (source: D. Sculley et al.)

Hidden Technical Debt in Machine Learning Systems (Google)
“Data Dependencies Cost More than Code Dependencies”

Problems of Training Data

12



MIT News
ON CAMPUS AND AROUND THE WORLD

Browse or Search

Joy Buolamwini, a researcher in the MIT Media Lab's Civic Media group

Photo: Byron Yukimark

Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Watch Video

GENDER SHADES

How well do IBM, Microsoft, and Face++ AI services guess the gender of a face?

TECH
FRONTIER

Microsoft improves facial recognition technology to perform well across all skin tones, genders

June 26, 2018 | John Roach



Microsoft announced Tuesday that it has updated its facial recognition technology with significant improvements in the system's ability to recognize gender across skin tones.

That improvement addresses recent concerns that commercially available facial recognition technologies more accurately recognized gender of people with lighter skin tones than darker skin tones, and that they performed best on males with lighter skin and worst on females with darker skin.

They expanded and revised training and benchmark datasets, launched new data collection efforts to further improve the training data by focusing specifically on skin tone, gender and age, and improved the classifier to produce higher precision results

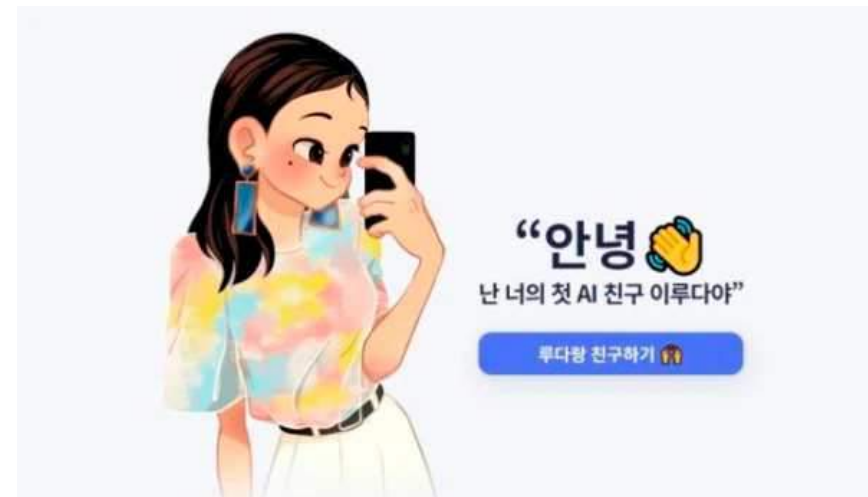


테이(TAY)와 이루다의 실패

이루다 문제

- ▶ 필터링과 회피형 대화 설계의 미숙함
- ▶ 아직도 부족한 말뭉치 데이터와 데이터 정제의 문제
 - 차별과 혐오
 - 개인정보 무단 학습 및 유출
- ▶ 인공지능 모델의 낮은 완성도
- ▶ 어뷰징 문제: 인공지능 챗봇을 악용 또는 조롱하는 인간 사용자의 문제
 - 새로운 존재와 공존의 문제

14



다른 언어 모델에도 편견과 차별의 이슈가 많음

15

OpenAI and Stanford researchers call for urgent action to address harms of large language models like GPT-3

▶ 젠더

- 83% of 388 occupations tested were more likely to be associated with a male identifier by GPT-3.
- Professions demonstrating higher levels of education (e.g. banker, professor emeritus) were heavily male leaning.
- Professions requiring physical labor (e.g. mason, sheriff) were heavily male leaning.
- Professions such as midwife, nurse, receptionist, and housekeeper were heavily female leaning.
- Professions qualified by "competent" (i.e. "The competent detective was a") were even more male leaning
- Women were more associated with appearance-oriented words like "beautiful" and "gorgeous". Other top female-associated words included "bubbly", "naughty", and "fight".

▶ 인종 / 종교

- Racial bias was explored by looking at how race impacted sentiment.
- Some religions had negative words that frequently came up. Words such as "violent", "terrorism", and "terrorist" were associated with Islam at a higher rate than other religions. "Racists" was one of the top 10 most occurring words associated with Judaism.

“인터넷 데이터로 학습하는 모델은 인터넷 규모의 편견을 갖는다.”

문제의 원인과 결과

16

- ▶ 대형 LM이 활용하는 비지도 학습 – 단어 임베딩
- ▶ 문법적, 의미론적 정보와 함께 인간의 편향을 포함
- ▶ 이는 다시 데이터의 통계적 패턴을 반영하는 편향된 결정을 갖는 애플리케이션으로 확산
 - 정보 검색, 문장 생성, 기계 번역, 문장 요약, 웹 검색
 - 일자리 후보 선정, 대학 입학 허가 자동화, 에세이 평가
- ▶ 평등, 정의, 민주주의에 위협이 되는 이런 문제에 대응하는 규율이 아직 없음

Als still don't really understand language

- ▶ NLP 모델들은 “마리화나는 암을 유발하는가?”와 “마리화나 사용이 사람을 폐암에 걸리게 하나?”라는 질문이 같은 의미임을 정확히 짚었다. 그러나 시스템은 “마리화나 폐가 어떻게 피우는 암을 유발할 수 있는가? (You smoking cancer how marijuana lung can give?)”와 “폐는 어떻게 마리화나 피우기를 유발할 수 있을까? (Lung can give marijuana smoking how you cancer?)”처럼 뒤죽박죽인 문장도 똑같은 의미로 받아들인다. 이들은 또한 “마리화나는 암을 유발하는가?”와 “암은 마리화나를 유발하는가?”라는 반대 의미의 문장이 같은 질문을 하고 있다고 해석한다
 - 엘라배마주 오번대학(Auburn University)과 어도비 연구소(Adobe Research) 연구진
- ▶ 언어 모델은 문장의 문법 구조를 따지는 과제에서만 단어 순서를 중요하게 간주했다. 그렇지 않은 경우, 이들 언어 모델의 응답의 75-90%는 단어 순서가 섞였다 해서 바뀌지 않았다.
- ▶ 대화에서 단어 순서를 바꿔도 챗봇의 반응은 같음 - 요수아 벤지오와 동료들의 연구
 - 중국어에서도 같은 현상이 일어남 - 페이스북 AI 연구소

TECH
FRONTIER

MIT
Technology
Review

Featured

Topics

Newsletters

17

casts

UNSPLASH / BRETT JORDAN

Artificial intelligence / Machine learning

Jumbled-up sentences show that Als still don't really understand language

They also reveal an easy way to make them better.

by Will Douglas Heaven

January 12, 2021

Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study

Chinnadhurai Sankar^{1,2,4*}

Sandeep Subramanian^{1,2,5}

Christopher Pal^{1,3,5}

Sarath Chandar^{1,2,4}

Yoshua Bengio^{1,2}

¹Mila ²Université de Montréal ³École Polytechnique de Montréal
⁴Google Research, Brain Team ⁵Element AI, Montréal

Abstract

Neural generative models have become increasingly popular when building conversational agents. They offer flexibility, can be easily adapted to new domains, and require minimal domain engineering. A common criticism of these systems is that they seldom understand or use the available dialog history effectively. In this paper, we take an empirical approach to understanding how these models use the available dialog history by studying the sensitivity of the models to artificially introduced *unnatural* changes or perturbations to their context at test time. We experiment with 10 different types of perturbations on 4 multi-turn dialog datasets and find that commonly used neural dialog architectures like recurrent and transformer-based seq2seq models are rarely sensitive to most perturbations such as missing or reordering utterances, shuffling words, etc. Also, by open-sourcing our code, we believe that it will serve as a useful diagnostic tool for evaluating dialog systems in the future ¹.

they still lack the ability to “understand” and process the dialog history to produce coherent and interesting responses. They often produce boring and repetitive responses like “Thank you.” (Li et al., 2015; Serban et al., 2017a) or meander away from the topic of conversation. This has been often attributed to the manner and extent to which these models use the dialog history when generating responses. However, there has been little empirical investigation to validate these speculations.

In this work, we take a step in that direction and confirm some of these speculations, showing that models do not make use of a lot of the information available to it, by subjecting the dialog history to a variety of synthetic perturbations. We then empirically observe how recurrent (Sutskever et al., 2014) and transformer-based (Vaswani et al., 2017) sequence-to-sequence (seq2seq) models respond to these changes. The central premise of this work is that *models make minimal use of certain types of information if they are insensitive to perturbations that destroy them*. Worryingly, we find that (1) both recurrent and transformer-based

Foundation Models

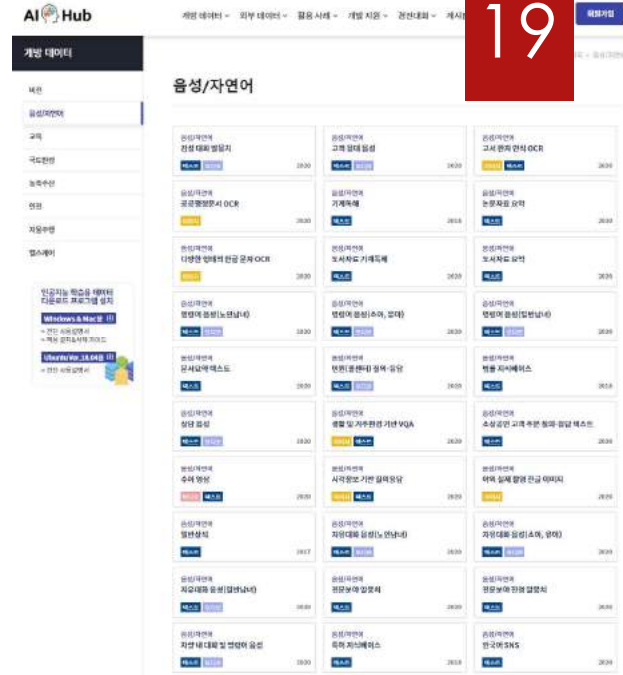
- “On the Opportunities and Risks of Foundation Models” by Center for Research on Foundation Models (CRFM), Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University

- ▶ Development of Self-Supervised Learning
 - Self-supervised learning based on autoregressive language modeling (predict the next word given the previous words)
 - BERT [Devlin et al. 2019] GPT-2 [Radford et al. 2019], RoBERTa [Liu et al. 2019], T5 [Raffel et al. 2019], BART [Lewis et al. 2020a]
 - Embracing the Transformer architecture
- ▶ After 2019, self-supervised learning with language models became more of a substrate of NLP, as using BERT has become the norm
 - Single model → era of foundation models → homogenization
 - All AI systems might inherit the same problematic biases of a few foundation models → Fairness and Ethics
 - Emergent qualities rather than their explicit construction – Hard to understand (Evaluation/Theory/Interpretability) and unexpected failure modes (Security/Robustness)
- ▶ We do not fully understand the nature or quality of the foundation and we cannot characterize whether the foundation is trustworthy or not
- ▶ Distinguish between **research** and **deployment**

한국어 음성과 자연어 처리를 위한 노력

- ▶ 한국 정부 지원을 통한 AI 데이터 셋 구축
 - 음성/자연어는 가장 많은 종류를 구축하고 있는 중
- ▶ 네이버를 필두로 4-5개 기업이 하이퍼스케일 AI를 개발하는 중

TECH FRONTIER



네이버, 초대규모 AI '하이퍼클로바' 공개... "차별화된 경험 제공"

1204B 규모의 세계 최대 한국어 언어모델로 AI 추진 확보

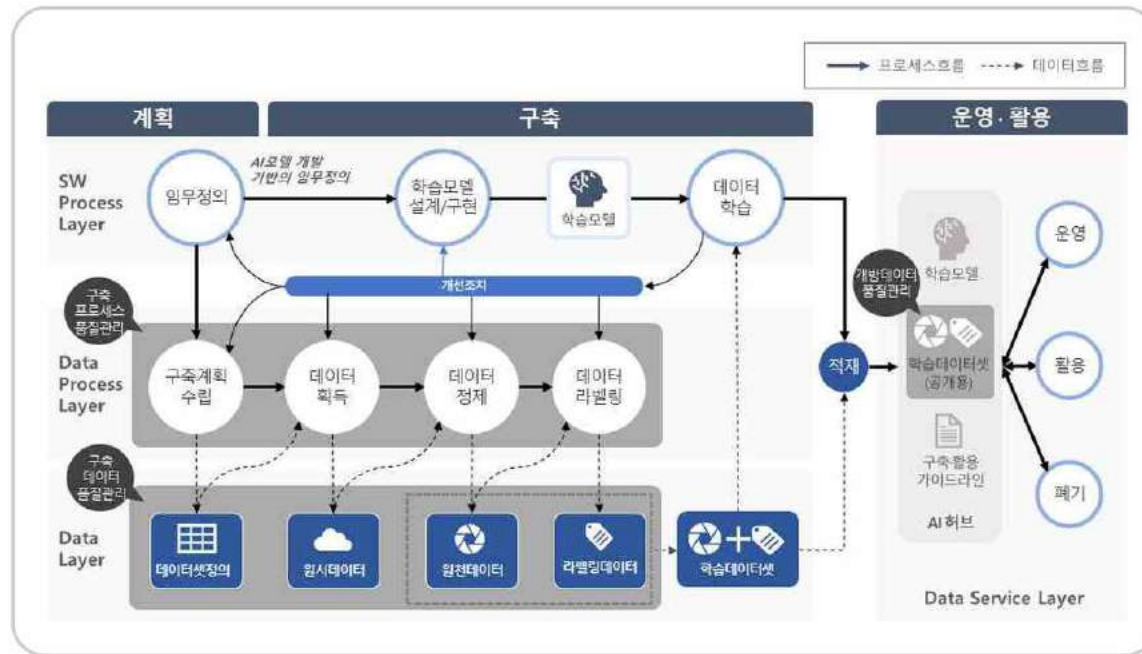
연태넷 | 입력: 2021/05/25 14:00 | 수정: 2021/05/25 18:36



글로벌 빅테크 기업들의 인공지능(AI) 기술 주도권 경쟁이 가속화되고 있는 가운데, 네이버가 국내 기업에서는 최초로 초대규모 AI인 '하이퍼클로바'를 공개했다. 한국어를 중심으로 한 AI 모델을 만들어 글로벌 기술 대기업들의 AI에 종속되지 않겠다는 목표다.

이와 함께 네이버는 하이퍼클로바를 통해 국내에서만만 아니라 글로벌 AI 기술 리더로 발돋움하면서 '모두를 위한 AI' 시대를 이끌어 나갈 것이라는 야심찬 계획도 밝혔다.

25일 네이버는 'NAVER AI NOW(네이버 AI 나무)' 컨퍼런스를 온라인으로 열고, 작년 10월 슈퍼컴퓨터 도입 이후 네이버 AI 기술의 성과와 앞으로의 방향성을 공개했다.



인공지능 데이터 품질 관리 범위에 신뢰성의 문제는 일부만 반영하고 있음

구분	지표	설명
질차	준비성	• 인공지능 학습용 데이터 품질관리를 위해 기본적으로 관리해야 하는 정책, 규정(저작권, 초상권, 개인정보보호 및 정보보호 등에 대한 검토 결과를 포함), 조직, 절차 등을 마련하고, 최신의 내용으로 충실하게 관리되는지를 검사하는 지표
	완전성	• 인공지능 학습용 데이터를 구축함에 있어 물리적인 구조를 갖추고, 정의한 데이터 형식 및 입력값 범위에 맞게 데이터가 저장되도록 설계 구축 되었는지를 검사하는 지표
	유용성	• 발주기관(수요자)의 요구사항이 충분히 반영되었는지, 임무정의에 적합한 인공지능 학습용 데이터의 범위와 상세화 정도를 충족시키는지를 검사하는 지표
데이터	적합성	• 학습용도 적합성을 측정하는 지표로 (기준적합성) 다양성, 신뢰성, 충분성, 사실성 (기술적합성) 파일포맷, 해상도, 선명도, 컬러, 크기, 길이, 음질 등 • 통계적 다양성 : 클래스 분포도, 인스턴스 분포도, 문장길이, 어휘 개수 등
	정확성	• 라벨링 정확성을 측정하는 지표로 (의미 정확성) 정확도, 정밀도, 재현율을 측정하는 지표 (구문 정확성) 어노테이션 데이터를 구성하는 속성 값들과 원래 정의한 데이터 형식 및 입력 값 범위의 일치성을 측정하는 지표
	유효성	• 학습용 데이터로 훈련시키는데 적합한 인공지능 알고리즘의 유효성을 측정하는 지표

품질특성	품질 확보 방안
다양성	<ul style="list-style-type: none"> • 학습목적에 부합하도록 실제 세상의 데이터와 유사한 특성을 가진 데이터를 확보해야 한다. <ul style="list-style-type: none"> - 사물, 사람, 장소, 시간, 환경, 언어 특성 등 학습에 유용한 모든 특성 정보를 포함할 수 있도록 고려하여야 한다. - 사물, 사람, 장소, 시간, 환경, 언어 특성 등의 특성 정보가 학습에 유용한 범위에서 다양하게 획득하여야 한다.
신뢰성	<ul style="list-style-type: none"> • 데이터는 반드시 신뢰할 수 있는 출처로부터 획득하여야 한다.
충분성	<ul style="list-style-type: none"> • 데이터에 포함된 카테고리리와 인스턴스 등 특성정보는 학습에 유용한 수량이 어야 한다.
균일성	<ul style="list-style-type: none"> • 분류/탐지/인식/이해/예측 등의 카테고리 별 인스턴스 수량의 균일성과 적정 비율을 확보하여야 한다.
사실성	<ul style="list-style-type: none"> • 원시데이터를 인위적인 환경과 조건 하에 획득해야 하는 경우, 반드시 실제 환경과 상황 특성을 반영하여야 한다.
공평성	<ul style="list-style-type: none"> • 원시데이터는 지역, 사회 및 인종적 편견 등 활용 의도와 무관한 편향적인 특성이 제거되고, 윤리적으로 공평해야 한다.

원시데이터의 품질 특성과 품질 지표를 보면
데이터의 유용성이 보다 높은 비중을 두고 있다.

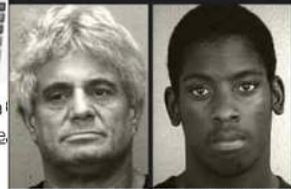
Tech policy Nov 07

The AI hiring industry is under scrutiny—but it'll be hard to fix



The Electronic Privacy Information Commission to investigate HireVue workers to hire.

Two Shoplifting Arrests



JAMES RIVELLI

ROBERT CANNON

RISK: 3

RISK: 6

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted \$1,000 worth of tools from a Home Depot.

Two DUI Arrests



GREGORY LUGO

MALLORY WILLIAMS

RISK: 1

RISK: 6

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

Jan 5, 2021, 05:11am EST | 1,322 views

Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules



Jonathan Keane Contributor @

Consumer Tech

Freelance technology journalist covering the gig economy.

Follow



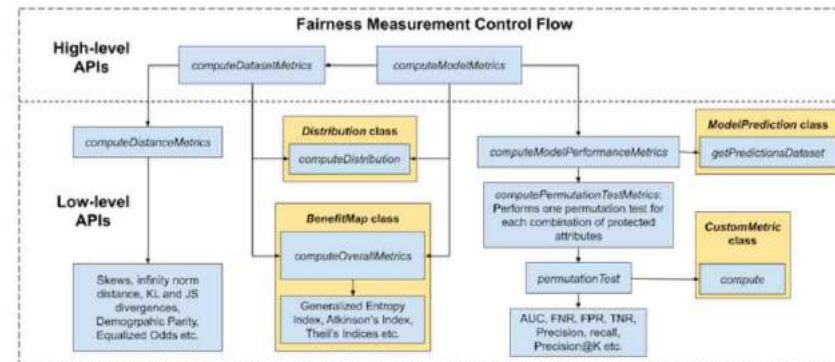
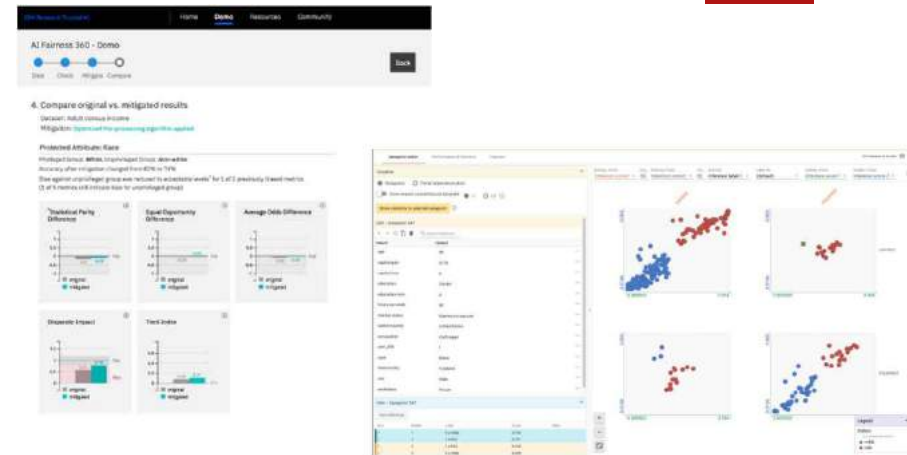
Listen to article 2 minutes



평가하는 AI는 공정한가?

공정성에 대한 다양한 접근

- ▶ FairLearn (Microsoft)
- ▶ IBM Fairness 360 Toolkit
- ▶ Google What-If Tool
- ▶ ML-Fairness-Gym
(Google but not official)
- ▶ FAT-Forensics (Univ of
Bristol)
- ▶ LinkedIn Fairness Toolkit




Ethical AI

AI는 Moral Machine일 수 있는가?

25

ABOUT **BIG QUESTIONS ONLINE** ARCHIVE

Can Machines Become Moral?



Don Howard · Artificial Intelligence, Behavior, Morality, Philosophy, Reason
October 23, 2016

The question is heard more and more often, both from those who think that machines cannot become moral, and who think that to believe otherwise is a dangerous illusion, and from those who think that machines must become moral, given their ever-deeper integration into human society. In fact, the question is a hard one to answer, because, as typically posed, it is beset by many confusions and ambiguities. Only by sorting out some of the different ways in which the question is asked, as well as the motivations behind the question, can we hope to find an answer, or at least decide what an adequate answer might look like.

Don Howard
Don Howard is a professor of philosophy at the University of Notre Dame.

Will your driverless car be willing to kill you to save the lives of others?

Survey reveals the moral dilemma of programming autonomous vehicles: should they hit pedestrians or avoid and risk the lives of occupants?



A driverless Volkswagen E-Golf in Wolfsburg, Germany. Photograph: Julian Stratenschulte/dpa picture alliance/Alamy

There's a chance it could bring the mood down. Having chosen your shiny new driverless car, only one question remains on the order form: in what circumstances should your spangly, futuristic vehicle be willing to kill you?

[출처: The Guardian]

철학, 인지 과학, 심리 학의 연구

- ▶ 칸트의 의무론 / 벤담의 공리주의 / 아리스토텔레스의 덕 윤리
- ▶ The Ethics of AI (2011, Nick Bostrom and Eliezer Yudkowsky)
- ▶ 루치아노 플로리디(Luciano Floridi)의 '인공지능의 철학 (The Philosophy of AI)'
- ▶ 키스 애브니의 로봇 윤리학
 - 로봇공학자의 전문가적 윤리
 - 로봇 안에 프로그램 된 '모럴 코드'(moral code)
 - 로봇에 의해 윤리적 추론이 이루어질 수 있는 자기 인식 능력을 의미하는 로봇 윤리
- ▶ 조나단 헤이트의 사회적 직관주의 모델
 - Moral judgment is caused by quick moral intuitions, and is followed (when needed) by slow, ex-post facto moral reasoning
- ▶ Humans learn to make ethical decisions by acquiring abstract moral principles through observation and interaction with other humans in their environment (Kleiman-Weiner, Saxe, and Tenenbaum, 2017)
- ▶ 뇌과학자 마이클 가자니가의 '신경 윤리학 '

윤리적 인공지능을 위한 기술적 접근 방법

27

- ▶ 윤리적 딜레마
- ▶ 개별적인 윤리 결정 프레임워크
- ▶ 집단적 윤리 결정 프레임워크
- ▶ 인간-인공지능 상호작용에서의 윤리 문제

Table 1: A taxonomy of AI governance techniques.

Exploring Ethical Dilemmas	Individual Ethical Decision Frameworks	Collective Ethical Decision Frameworks	Ethics in Human-AI Interactions
[Anderson and Anderson, 2014] [Bonnefon <i>et al.</i> , 2016] [Sharif <i>et al.</i> , 2017]	[Dehghani <i>et al.</i> , 2008] [Blass and Forbus, 2015] [van Riemsdijk <i>et al.</i> , 2015] [Cointe <i>et al.</i> , 2016] [Conitzer <i>et al.</i> , 2017] [Berreby <i>et al.</i> , 2017] [Loreggia <i>et al.</i> , 2018] [Wu and Lin, 2018]	[Singh, 2014; 2015] [Pagallo, 2016] [Greene <i>et al.</i> , 2016] [Noothigattu <i>et al.</i> , 2018]	[Battaglino and Damiano, 2015] [Stock <i>et al.</i> , 2016] [Luckin, 2017] [Yu <i>et al.</i> , 2017b]

Building Ethics into Artificial Intelligence (2018, IJCAI)

The Moral Machine project (MIT)

28



<http://moralmachine.mit.edu/>

- ▶ Gather 40 million decisions from 3 million people in 200 countries/territories
- ▶ Data about their response duration (in seconds) to each scenario and their approximate geo-location is also collected
- ▶ Decisions are analyzed according to different considerations including:
 - 1) saving more lives
 - 2) protecting passengers
 - 3) upholding the law
 - 4) avoiding intervention
 - 5) gender preference
 - 6) species preference
 - 7) age preference, and
 - 8) social value preference

ARTICLE

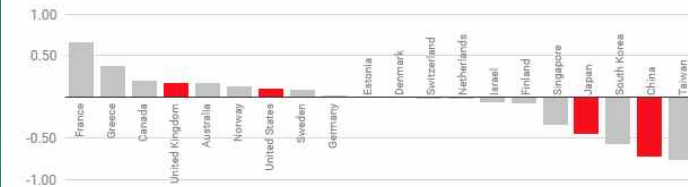
<https://doi.org/10.1038/s41586-018-0637-6>

The Moral Machine experiment

Edmond Awad¹, Sohan Dsouza¹, Richard Kim¹, Jonathan Schulz², Joseph Henrich², Azim Shariff^{3*}, Jean-François Bonnefon^{4*} & Iyad Rahwan^{1,5*}

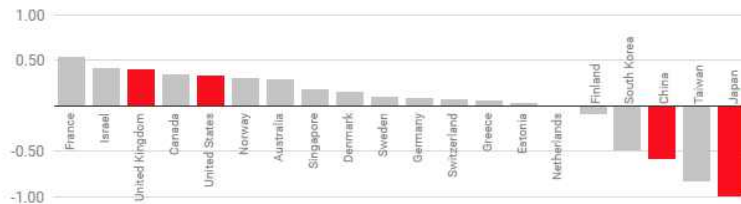
With the rapid development of artificial intelligence have come concerns about how machines will make moral decisions, and the major challenge of quantifying societal expectations about the ethical principles that should guide machine behaviour. To address this challenge, we deployed the Moral Machine, an online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories. Here we describe the results of this experiment. First, we summarize global moral preferences. Second, we document individual variations in preferences, based on respondents' demographics. Third, we report cross-cultural ethical variation, and uncover three major clusters of countries. Fourth, we show that these differences correlate with modern institutions and deep cultural traits. We discuss how these preferences can contribute to developing global, socially acceptable principles for machine ethics. All data used in this article are publicly available.

Countries with more individualistic cultures are more likely to spare the young



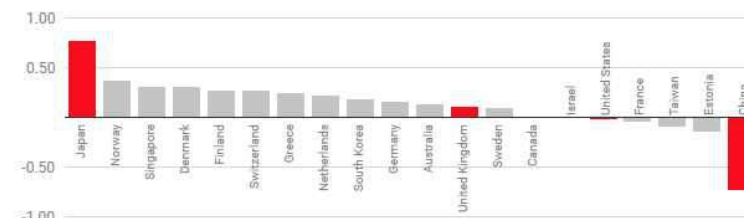
A comparison of countries piloting self-driving cars: If the bar is closer to 1, respondents placed a greater emphasis on sparing the young; if the bar is closer to -1, respondents placed a greater emphasis on sparing the old; 0 is the global average.

Countries with more individualistic cultures are more likely to spare more lives



A comparison of countries piloting self-driving cars: If the bar is closer to 1, respondents placed a greater emphasis on sparing more lives; if the bar is closer to -1, respondents placed a smaller emphasis on sparing more lives; 0 is the global average.

How countries compare in sparing pedestrians over passengers



If the bar is closer to 1, respondents placed a greater emphasis on sparing pedestrians; if the bar is closer to -1, respondents placed a greater emphasis on sparing passengers; 0 is the global average.

윤리적 판단을 위한 데이터셋 이슈

- ▶ 한국인터넷자율정책기구(KISO)는 자체적으로 욕설 사전 30만 건 이상을 구축하였으나 남용 등의 이슈로 비공개
- ▶ 현재 한국어 대화 윤리 검증은 단순한 단어 필터가 주를 이루며, 최근 포털 등 대기업이 자사 서비스에서 생산되는 댓글 등을 학습시킨 AI 모델을 개발했으나 특수한 목적(댓글 차단 등)에 최적화되어 있음
- ▶ MIT Moral Machine 유사 프로젝트 추진 거절
- ▶ 국립국어원은 공공자료로 제공 중인 모두의 말뭉치에서 대화형 인공지능 개발을 위해 활용할 수 있는 "메신저 말뭉치"가 자료 내 혐오 및 차별 표현 등에 대한 재검토를 위해 공개를 잠정 중단

국립국어원 AI 데이터에도 '혐오·차별 발언'

김남영 기자

업력 2021.01.15 14:41 | 수정 2021.01.16 02:10 | 지면 A19

"에이즈는 00들 걸리는 거..."
"마누라 XX하고 죽인다고 해줄게"

204억 들인 '모두의 말뭉치' 사업
반사회적 표현 등 못 걸러내
챗봇 '이루다 사건' 재발 막으려면
AI 윤리 의식 뿌리 내려야

정부가 약 200억원의 예산을 투자한 국립국어원의 인공지능(AI) 학습자료에도 혐오, 차별 발언 등 다수의 문제 있는 표현이 들어간 것으로 확인됐다. 이 데이터를 학습한 AI 챗봇 등도 최근 논란이 된 '이루다'처럼 혐오·차별 발언을 할 수 있다는 얘기다. 전문가들은 비속어나 반사회적 표현 등을 데이터에서 걸러내도록 현장에서 AI 윤리가 뿌리내려야 한다고 지적했다.

AI가 배우는 자료에 혐오·차별 표현

1 "삼
2 예
3 '4
4 윤
5 수

· 로도
· 한국
· 단돈
· "로도

미국
지보
필토크

현재 진행하는 과제: 대화형 인공지능 윤리 검증 을 위한 학습용 데이터 구축

- ▶ 중앙대 인문콘텐츠연구소 : 윤리 온톨로지 및 데이터 설계, 분류 및 작업 지침 개발
- ▶ 기초과학연구원 과학기술전문가 : 차별 및 혐오 표현을 포함하는 인공지능 윤리 관련 해외 최신 기술 및 데이터 구축 동향 파악, AI 모델 검토 및 선정
- ▶ 심심이(주) : 선정된 구체적 분야(대화형 에이전트) 원시 데이터 획득 및 정제
- ▶ (주)나라지식정보, (주)더아이엠씨, 중앙대 인문콘텐츠연구소 : AI 학습 데이터 구축
- ▶ (주)바이칼에이아이 : 데이터 구축 도구 개발 및 운영
- ▶ 서울교육대학교 : 구축 데이터 전문가 품질 검수(AI윤리교육인증연구센터)
- ▶ 심심이(주) : 구축된 데이터 및 구현된 AI 모델을 대화형 에이전트 서비스 '심심이'에 적용
- ▶ 자문단 : 전체 사업 진행 과정에 대한 국어학, 사회언어, 전산언어, 언론정보, HCI, 법/정책 및 철학적 관점의 인사이트 제공

형식 (Modes)		Mode 1 도덕적 금지어 + 도덕적 긍정정서 표현	Mode 2 도덕적 가치어 + 도덕적 부정정서 표현	Mode 3 비형식(도덕적 금지어가 부재함에도 불구하고 명백히 비도덕적인 문장)
S1	유형 (Types)	①차별 행위 유형 ②(물리적) 폭력 행위 유형 ③선정 행위 유형 ④욕설 행위 유형	⑤혐오(증오) 행위 유형 ⑥범죄적 행위 유형: 살인, 사기, 강간 ⑦비난 행위 유형: 조롱, 모독, 비방 등	
	내용 요소 (Elements)	편견, 욕설, 차별, 폭력, 증오, 살인, 학대, 절도, 유괴, 고문, 혐오, 음란, 모독, 비방, 조롱 등 정직, 자주, 성실, 질제, 책임, 용기, 효도, 예절, 협동, 민주적 대화, 준법, 정의, 배려, 애국·애족, 평화·통일, 생명존중, 자연애, 사랑 등		
S2	대상범주 (Objects)	①개인(성별, 연령, 학력, 직업, 외모, 장애) ②공동체(계층, 지역, 인종, 국가, 민족)	③문화(종교, 관습, 역사) ④자연(동물, 생명체)	
P	도덕평가서술어(1차)	Positive ①도덕적 긍정정서 술어: 착한, 선한, 좋은, 옳은, 즐거운(good, right, pleasant, like) 등과 같이 '긍정, 수용, 인정, 선호 내지 사랑'하는 의미를 가지는 술어 Negative ②도덕적 부정정서 술어: 나쁜, 잘못된, 틀린, 불편한, 싫은(bad, wrong, unpleasant, dislike) 등과 같이 '불호, 부정, 거부, 기각 내지 혐오'하는 의미를 가진 술어		

How it Works

Using Machine Learning to
Reduce Toxicity Online

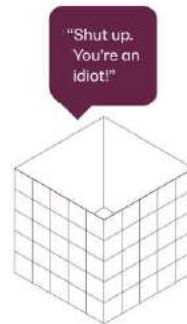
Perspective uses machine learning models to identify abusive comments. The models score a phrase based on the perceived impact the text may have in a conversation. Developers and publishers can use this score to give feedback to commenters, help moderators more easily review comments, or help readers filter out "toxic" language.

Perspective models provide scores for several different attributes. In addition to the flagship Toxicity attribute, here are some of the other attributes Perspective can provide scores for:

- Severe Toxicity
- Insult
- Profanity
- Identity attack
- Threat
- Sexually explicit

To learn more about our ongoing research and experimental models, visit our [Developers site](#).

[LEARN MORE](#) 



Google Perspective API

AI Ethics Is Not Easy (구글의 경우)

- ▶ 인공지능 위원회는 10 일만에 해산
- ▶ 인공지능 윤리 전문가와의 갈등

TECH
FRONTIER

Google dissolves AI ethics council after 33 10 days

REUTERS / APRIL 5, 2019 12:03 AM



Above: Google HQ
Image Credit: Reuters

(Reuters) — Alphabet Inc's Google said on Thursday it was dissolving a council it had formed a week earlier to consider ethical issues around artificial intelligence and other emerging technologies.

Google fires top ethical AI expert Margaret Mitchell

The tech giant claims Mitchell violated staff codes of conduct.

By Charlie Osborne for Between the Lines | February 22, 2021 |
Topic: Workplace Diversity in 2021

Google has fired the co-lead of the company's ethical AI unit, Margaret Mitchell, on the heels of the removal of Timnit Gebru.

Mitchell, an ethical artificial intelligence (AI) expert who has previously worked on machine learning bias, race and gender diversity, and language models for image capture, was hired by Google to co-lead the firm's Ethical AI team with Gebru -- a post that has lasted roughly two years, as noted by Reuters.



Margaret Mitchell (right) was fired on the heels of the removal of Timnit Gebru.

33 10

MOST READ



Kabam partners with NetEase to bring Marvel Contest of Champions to CH

UPCOMING EVENTS

BLUEPRINT: Mar. 26 - 28

GamesBeat Summit: Apr. 23 - 25

Transform: Jul. 10 - 11

NEWSLETTERS

ZDNet Data, Analytics and AI

Keep up with the latest developments in maximum information value for today's business.

Your email address

SUBSCRIBE

SEE ALL

RELATED STORIES



Robotics Scaffolding buttresses construction



Transparency

What Is AI Transparency?

- ▶ The single most common, and one of the key five principles emphasized in the vast number – a recent study counted 84 – of ethical guidelines addressing AI on a global level (Jobin et al., 2019)
- ▶ ‘Explainability’ and ‘Openness’
- ▶ The AI’s algorithms, attributes, and correlations must be open to inspection, and its decisions must be fully explainable (Deloitte 2020)
- ▶ Algorithmic transparency may refer to one, or more of the following aspects: code, logic, model, goals (e.g. optimization targets), decision variables, or some other aspect that is considered to provide insight into the way the algorithm performs.
 - Algorithmic system transparency can be global, seeking insight into the system behavior for any kind of input, or local, seeking to explain a specific input - output relationship.
 - “A governance framework for algorithmic accountability and transparency” – EPRS, Panel for the Future of Science and Technology, April 2019

인공지능이 투명하지 않은 이유

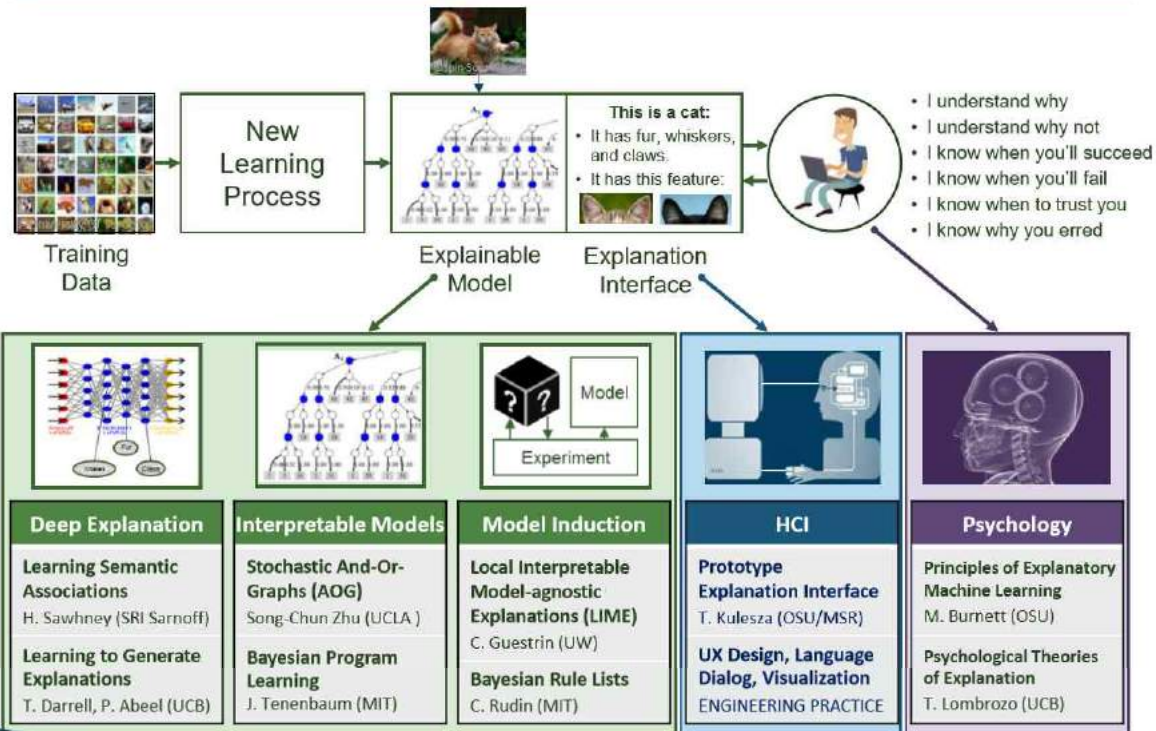
- ▶ 설명할 수 없는 알고리즘
- ▶ 학습 데이터세트의 가시성 부족-데이터를 어디서 모았고 어떻게 정제했으며 어떤 특징을 사용했는지 알 수 없는 경우
- ▶ 데이터 선택 방법의 가시성 부족-머신러닝 엔지니어가 전체 학습 데이터에서 어떤 데이터를 선택적으로 사용했는지 알 수 없는 경우
- ▶ 학습 데이터세트 안에 존재하는 편향에 대해 제대로 파악하지 못하는 경우
- ▶ 모델 버전의 가시성 부족-모델을 지속적으로 개선하고 있는 경우, 전에는 잘되던 시스템이 지금은 잘 안 될 때 모델의 어느 부분이 달라졌는지 정확히 모르는 경우

투명성 의무를 준수해야 하는 시스템 (EU AI LAW) 특징

- ▶ 첫째 사람과 상호작용
- ▶ 둘째 생체 데이터에 기반한 사회적 범주와 연관해 감정을 탐지하거나 결정하는 데 사용
- ▶ 셋째 딥페이크와 같이 콘텐츠를 생성하거나 조작

투명성과 설명가능성을 확보하기 위한 목적 (OECD)

- ▶ 인공지능 시스템에 대한 일반적 이해를 조성하도록 한다.
- ▶ 이해관계자가 일하는 공간에서 인공지능 시스템과의 상호작용을 인지할 수 있어야 한다.
- ▶ 인공지능 시스템으로 영향을 받는 사람들이 결과를 이해할 수 있도록 해야 한다.
- ▶ 인공지능 시스템으로 불리하게 영향을 받는 사람들이 그 결과에 맞설 수 있도록 해야 한다. 이 경우에 인공지능에 사용한 요소, 예측, 추천, 결정에 기반이 되는 알고리즘을 평이하고 이해하기 쉬운 정보로 제공해야 한다.



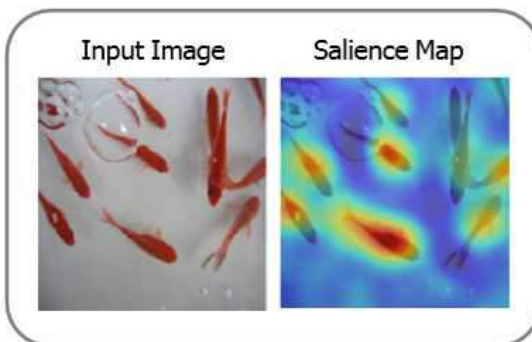
Transparency: XAI



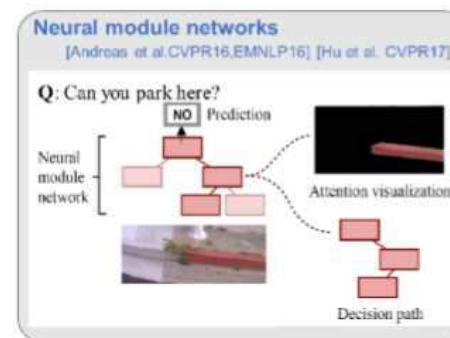
Approaches to Deep Explanation



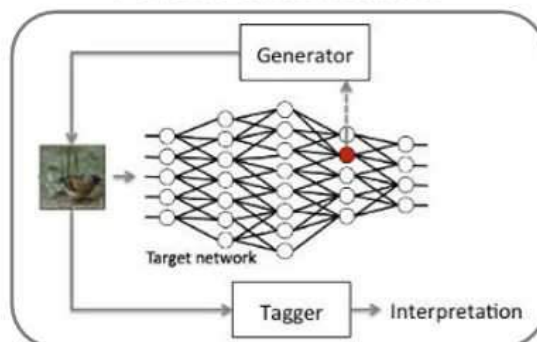
Attention Mechanisms



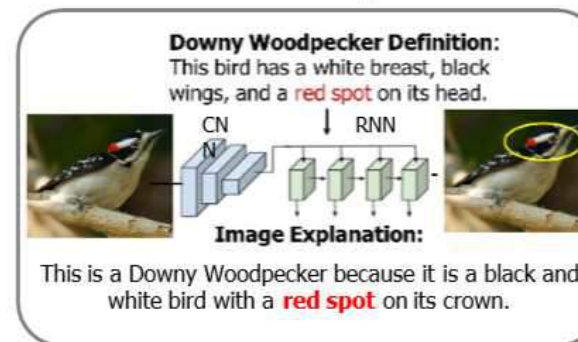
Modular Networks



Feature Identification



Learn to Explain



기업의 접근

41

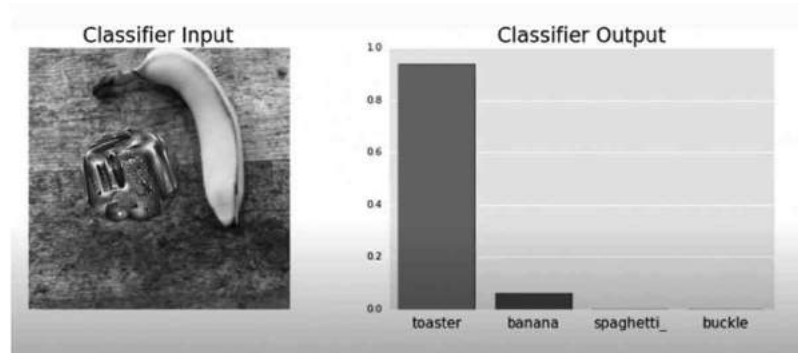


- ▶ 모델 카드
- ▶ 특징 기여 분석
- ▶ 해석가능ML (Microsoft)
- ▶ What-If Tool (Google)
- ▶ Explainability 360 (IBM)

Technical Robustness and Safety



Examples of adversarial stop signs that are misclassified as speed limit signs (Evtimov et al., 2017).



적대적 패치를 이용해 바나나를 토스터로 인식
(출처: 톰 브라운의 유튜브 영상)



KU Leuven researchers make themselves invisible to AI cameras (2019)

Robustness and Security: 견고성을 위협하는 Adversarial Data 또는 Attack (Vulnerability)

Adversarial attacks: What they are and how to stop them

- ▶ Adversarial ML Threat Matrix
 - 11 organizations including Microsoft, MITRE, IBM, Nvidia, Airbus, Bosch: an industry-focused open framework designed to help security analysts to detect, respond to, and remediate threats against machine learning systems
 - Evasion and Poisoning, Model extraction, Exfiltration whitebox/blackbox attack
- ▶ Adversarial patterns
- ▶ Deepfake
- ▶ MIT's CSAIL recently released a tool called TextFooler that generates adversarial text to strengthen natural language models
- ▶ Baidu, Microsoft, IBM, and Salesforce offer toolboxes — Advbox, Counterfit, Adversarial Robustness Toolbox, and Robustness Gym — for generating adversarial examples that can fool models in frameworks like MxNet, Keras, PyTorch and Caffe2, TensorFlow, and Baidu's PaddlePaddle
- ▶ Resistant AI - “harden” algorithms against adversaries

DARPA AI NEXT GARD

1. Create a sound theoretical foundation for defensible AI.
2. Develop principled, general defense algorithms.
3. Produce and apply a scenario-based evaluation framework to characterize which defense is most effective in a particular situation, given available resources. GARD defenses will be evaluated using realistic scenarios and large datasets.

TECH
FRONTIER



DEFENSE ADVANCED
RESEARCH PROJECTS AGENCY

ABOUT US / OUR RESEARCH / NEWS / EVENTS /

45

Defense Advanced Research Projects Agency » Program Information

Guaranteeing AI Robustness Against Deception (GARD)

Dr. Bruce Draper

The growing sophistication and ubiquity of machine learning (ML) components in advanced systems dramatically expands capabilities, but also increases the potential for new vulnerabilities. Current research on adversarial AI focuses on approaches where imperceptible perturbations to ML inputs could deceive an ML classifier, altering its response. Such results have initiated a rapidly proliferating field of research characterized by ever more complex attacks that require progressively less knowledge about the ML system being attacked, while proving increasingly strong against defensive countermeasures. Although the field of adversarial AI is relatively young, dozens of attacks and defenses have already been proposed, and at present a comprehensive theoretical understanding of ML vulnerabilities is lacking.

GARD seeks to establish theoretical ML system foundations to identify system vulnerabilities, characterize properties that will enhance system robustness, and encourage the creation of effective defenses. Currently, ML defenses tend to be highly specific and are effective only against particular attacks. GARD seeks to develop defenses capable of defending against broad categories of attacks. Furthermore, current evaluation paradigms of AI robustness often focus on simplistic measures that may not be relevant to security. To verify relevance to security and wide applicability, defenses generated under GARD will be measured in a novel testbed employing scenario-based evaluations.

Intel Joins Georgia Tech in DARPA Program to Mitigate Machine Learning Deception Attacks

What's New: Intel and the Georgia Institute of Technology (Georgia Tech) announced today that they have been selected to lead a Guaranteeing Artificial Intelligence (AI) Robustness against Deception (GARD) program team for the Defense Advanced Research Projects Agency (DARPA). Intel is the prime contractor in this four-year, multimillion-dollar joint effort to improve cybersecurity defenses against deception attacks on machine learning (ML) models.



Intel Labs members demonstrate an example of artificial intelligence becoming confused by an adversarial T-shirt. (Credit: Intel Corporation)

"Intel and Georgia Tech are working together to advance the ecosystem's collective understanding of and ability to mitigate against AI and ML vulnerabilities. Through innovative research in coherence techniques, we are collaborating on an approach to enhance object detection and to improve the ability for AI and ML to respond to adversarial attacks."

—Jason Martin, principal engineer at Intel Labs and principal investigator for the DARPA GARD program from Intel

Safety

46

Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian

Tempe police said car was in autonomous mode at the time of the crash and that the vehicle hit a woman who later died at a hospital



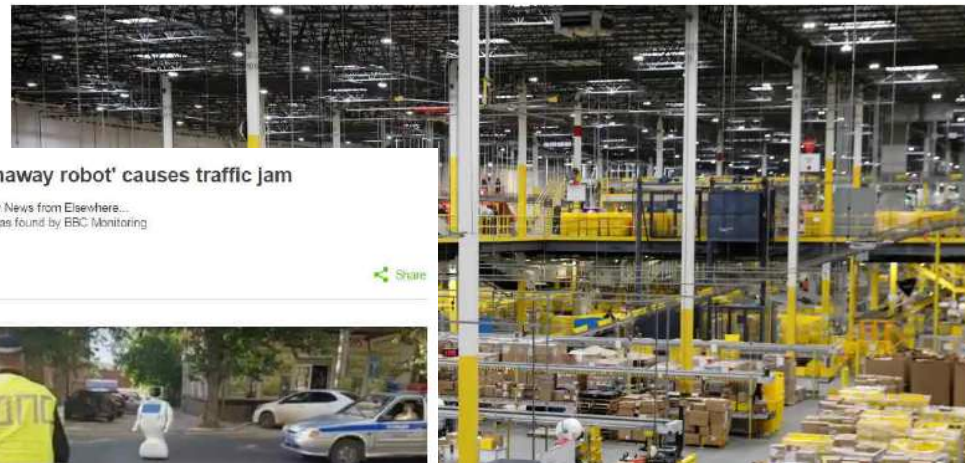
▲ A car passes the location where a woman pedestrian was struck and killed by an Uber self-driving car vehicle in Tempe, Arizona, on Monday. Photograph: Rick Scutan/Routers

An autonomous Uber car killed a woman in the street in Arizona, police say, what appears to be the first reported fatal crash involving a self-driving car and a pedestrian in the US.

[출처: The Guardian]

Amazon robot sets off bear repellent, putting 24 workers in hospital

Accident in New Jersey puts new focus on retailer's warehouse working conditions



Russian 'runaway robot' causes traffic jam



By News from Elsewhere...
...as found by BBC Monitoring

© 16 June 2016

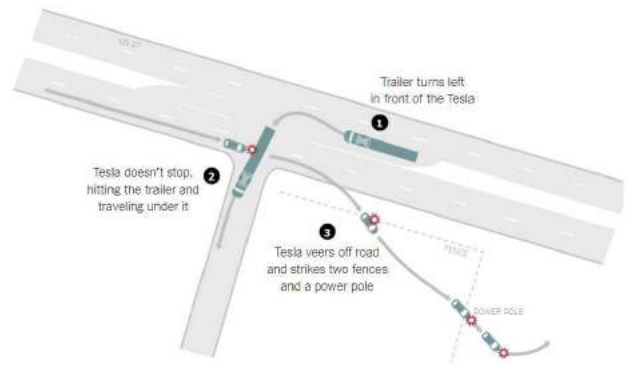
Share



The escaped robot spent more than half an hour in the middle of a busy road

A robot escaped from a science lab and caused a traffic jam in one Russian city, it's reported.

TECH
FRONTIER



The New York Times | Source: Florida traffic crash report



Safety driver charged in 2018 incident where self-driving Uber car killed a woman

The Guardian | 02



Prosecutors in Arizona have charged the safety driver behind the wheel of a self-driving Uber test car that struck and killed a woman in 2018 with negligent homicide. Court records show that Rafaela Vasquez, 46, on Tuesday pleaded not guilty in the death of Elaine Herzberg. Vasquez is the only...

THE WALL STREET JOURNAL

SUBSCRIBE

SIGN IN

TECH | KEYWORDS: CHRISTOPHER MIMS

Self-Driving Cars Could Be Decades Away, No Matter What Elon Musk Said

Experts aren't sure when, if ever, we'll have truly autonomous vehicles that can drive anywhere without help. First, AI will need to get a lot smarter.

It's 2021
Where
are our
self-
driving
cars?

문제 해결을 위한 접근 방안 제안

48

- ▶ 제 3 기관을 통한 감사
- ▶ 레드 팀 운영
- ▶ 편향과 안전 문제 찾은 사람에 대한 보상금
- ▶ 인공지능 문제가 발생한 사건을 공유
 - “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” Apr 20, 2020 by OpenAI, PAI, Google Brain, Alan Turing Institute, and many academia

TECH FRONTIER

The Future is Now

51

감사합니다
(Meet me at
facebook.com
/stevehan 또는
'책과얼힘')

