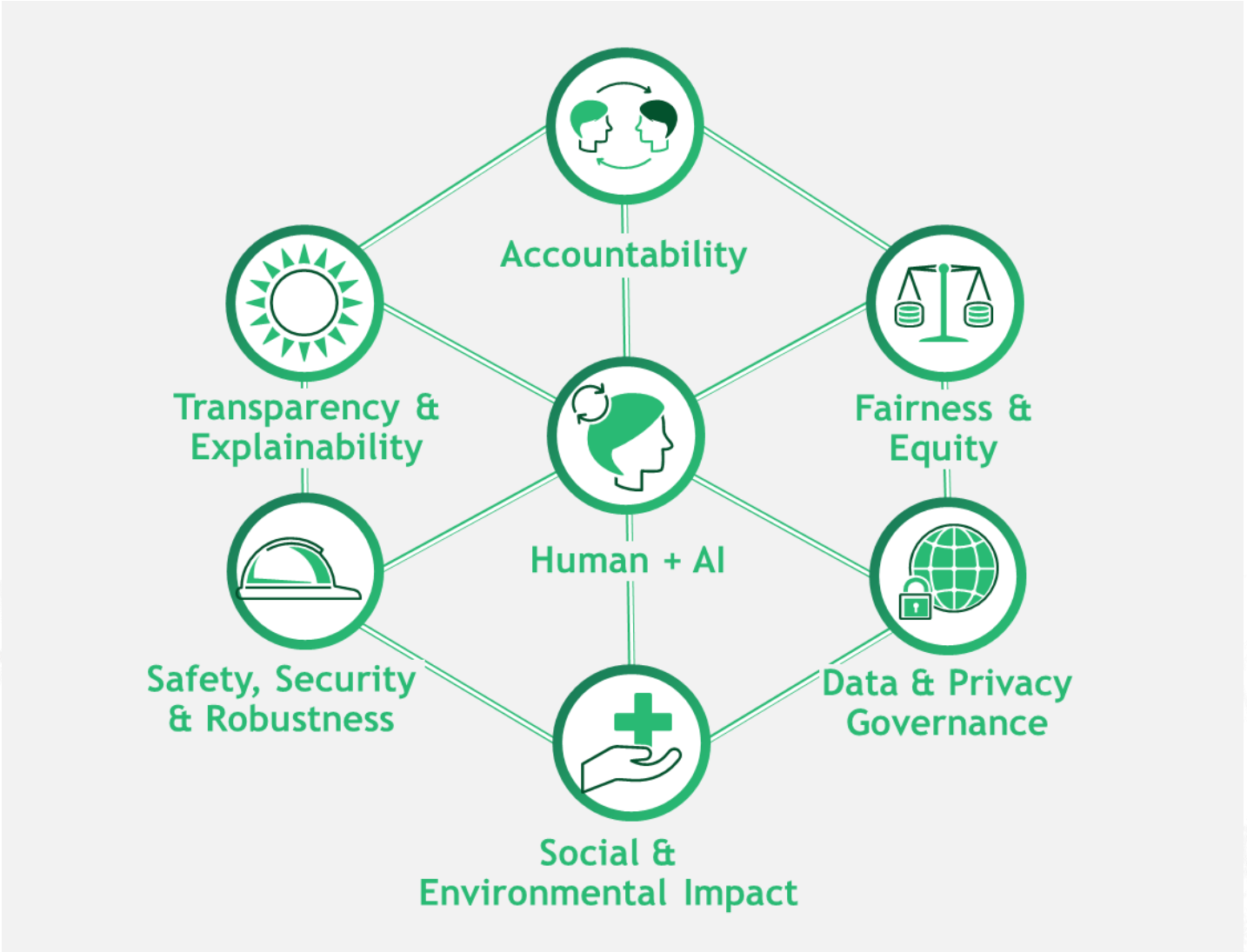
Responsible AI

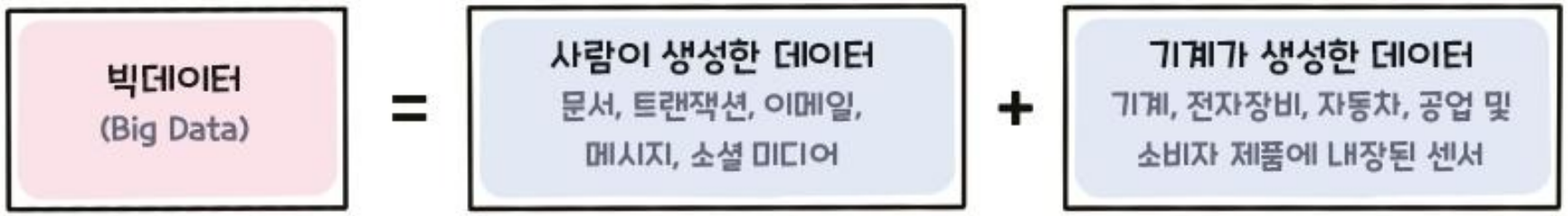
Responsible AI



Hyperscale Data

❖ Concepts

- Big Data
  - Data generated in a ***digital environment***
  - Big data containing ***text and video data*** as well as ***numerical data***, is massive compared to data produced in analog environments in the past



Hyperscale Data

❖ Concepts

- Difference Between Big Data and Traditional General Data
  - Big data means collected from various methods, sources, and environment
  - Big data is big enough to require a computer system for parallel processing
  - Big data creates value for business or research
  - Validity(타당성) and reliability(신뢰성) must be secured to ensure the value created by big data

구분	일반 데이터	빅데이터
데이터의 원천	내부로부터 수집	외부로부터 수집
데이터의 형태	정형 데이터가 대부분	비정형 데이터
분석 방법	모델링	인공지능
분석 환경	기업 내에 구축된 데이터웨어하우스	클라우드

Public Issues for Hyperscale Data

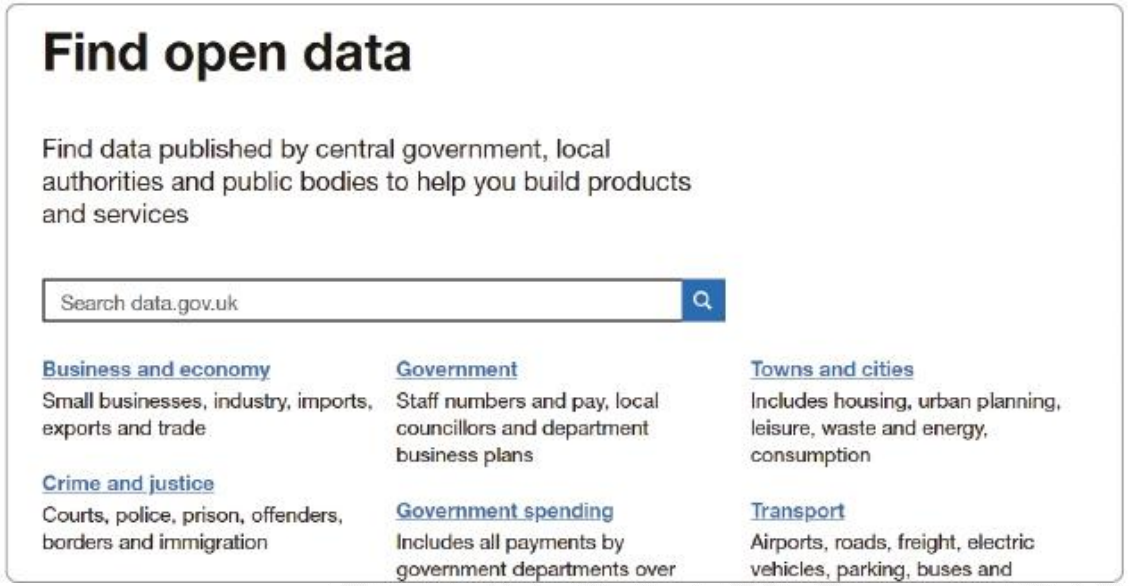
- Necessity of public data
  - The ***short-term way to achieve the effects of big data*** is to utilize government-owned data
  - If the government opens up big data, the private sector will be able to develop services without much effort

Public Issues for Hyperscale Data

- Examples of public data
  - US launches data portal site (data.gov) in 2009 to open public data
    - The U.S. government's data openness ranges from agriculture, environment and energy sectors to budgets, expenditures, contracts, and civil servants' salaries for financial transparency

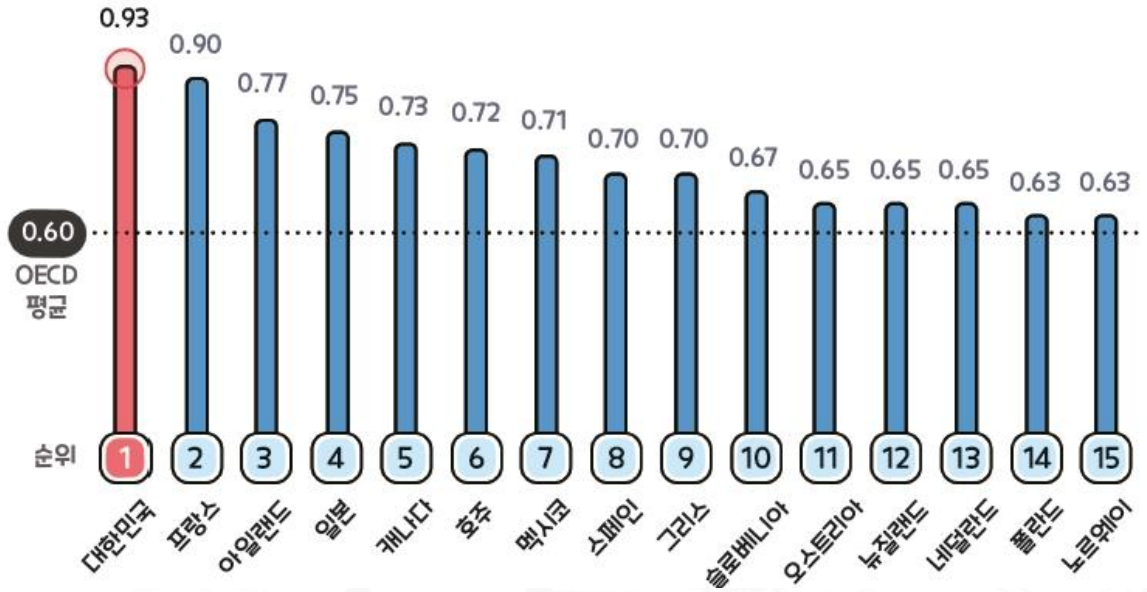
Public Issues for Hyperscale Data

- Examples of public data
  - UK began opening public data in 2010 when Prime Minister Gordon Brown invited Tim Berners-Lee, the founder of web and linked data, to create a data portal site (data.gov.uk)
    - Since 2018, it has been reorganized under the name of "Find Open Data"



Public Issues for Hyperscale Data

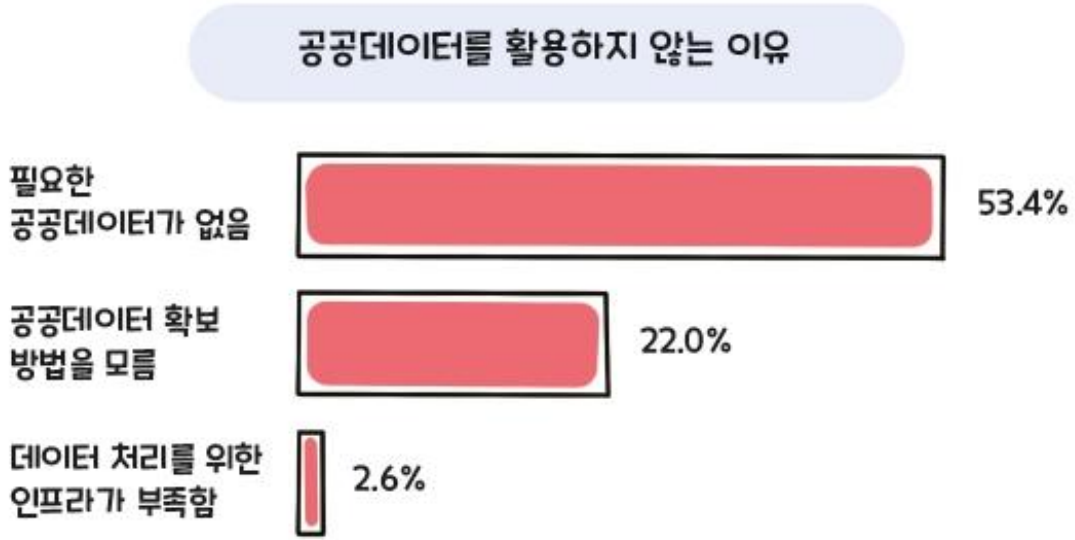
- Examples of public data
  - South Korea ranked ***first in the "2019 OECD White Paper" released by the OECD*** with 0.93 points in the public data opening index



Public Issues for Hyperscale Data

❖ Utilization of public data

- While Korea's public data openness index is high, the reason why the ***actual public data utilization rate is not so high*** is that ***<u>the utilization is low because the necessary data cannot be found</u>***



Public Issues for Hyperscale Data

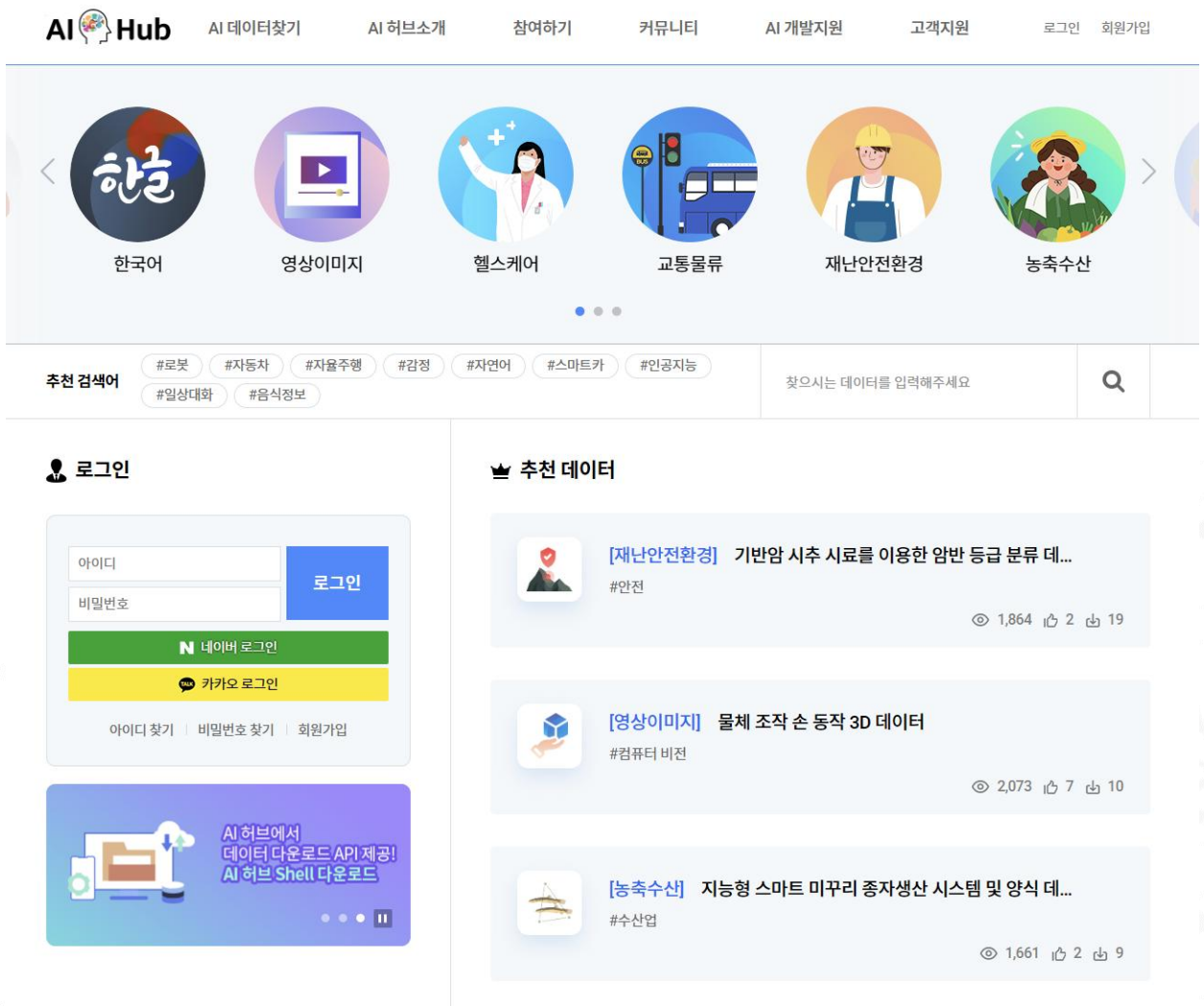
❖ Utilization of public data

분야	개방 예정인 공공데이터
자율주행(11개)	정밀도로지도, 주행환경 인식센서 융합정보, 자율주행 딥러닝 학습 정보 등
스마트시티(6개)	스마트 전력거래, 디지털 트윈 정보, 세종시 스마트에너지 정보 등
헬스케어(8개)	해부학 그림 및 의료행위 그림 정보, 한의약 전주기 정보, 식중독균 유산균 유전체 정보 등
금융정보(5개)	상장사 공시주식 정보, 비상장사 공시 재무제표, 주택저당채권 정보 등
생활환경(7개)	굴뚝 대기오염물질 정보, 산림 미세먼지 정보, 산업부문 온실가스 배출정보 등
재난안전(9개)	구조구급활동 정보, 산사태 정보, 안전·취약시설물관리 정보, 국가화재 정보 등

Public Issues for Hyperscale Data

❖ Utilization of public data

https://www.aihub.or.kr/



Technology Trends for AI

❖ Transparency(투명성) and **Reliability**(신뢰성)

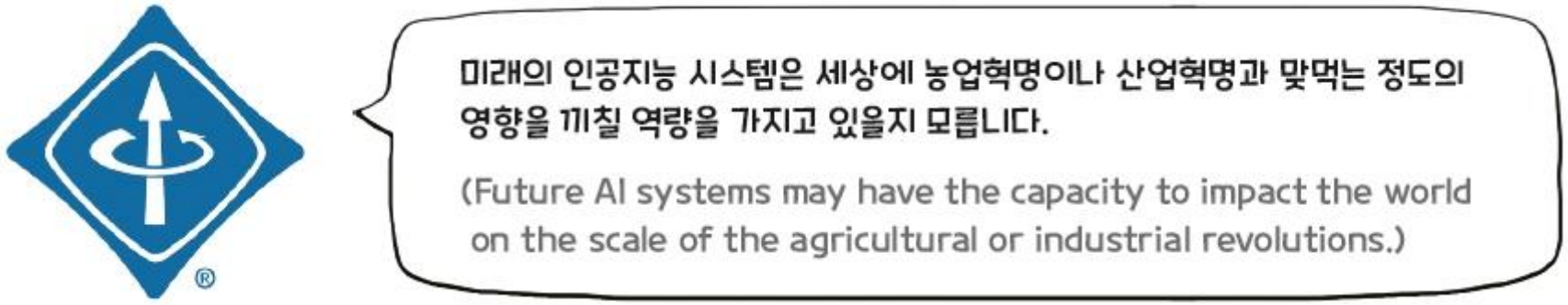
- AI causes ***unexpected errors*** due to the complexity of biased data or algorithms
- The problem is that current AI is ***a black box system*** that cannot explain the reason for its decision
- Unless the transparency of the learning process and results of AI is ***verified, reliability is of course a problem that must be raised***

Technology Trends for AI

- Transparency and **Reliability**
  - ***Various policies and guidelines for AI reliability verification are being announced***, but they are mainly related to follow-up measures, so they cannot be a fundamental prevention
  - Verification of data, the beginning of artificial intelligence, can be said to be the beginning of guaranteeing the reliability of artificial intelligence

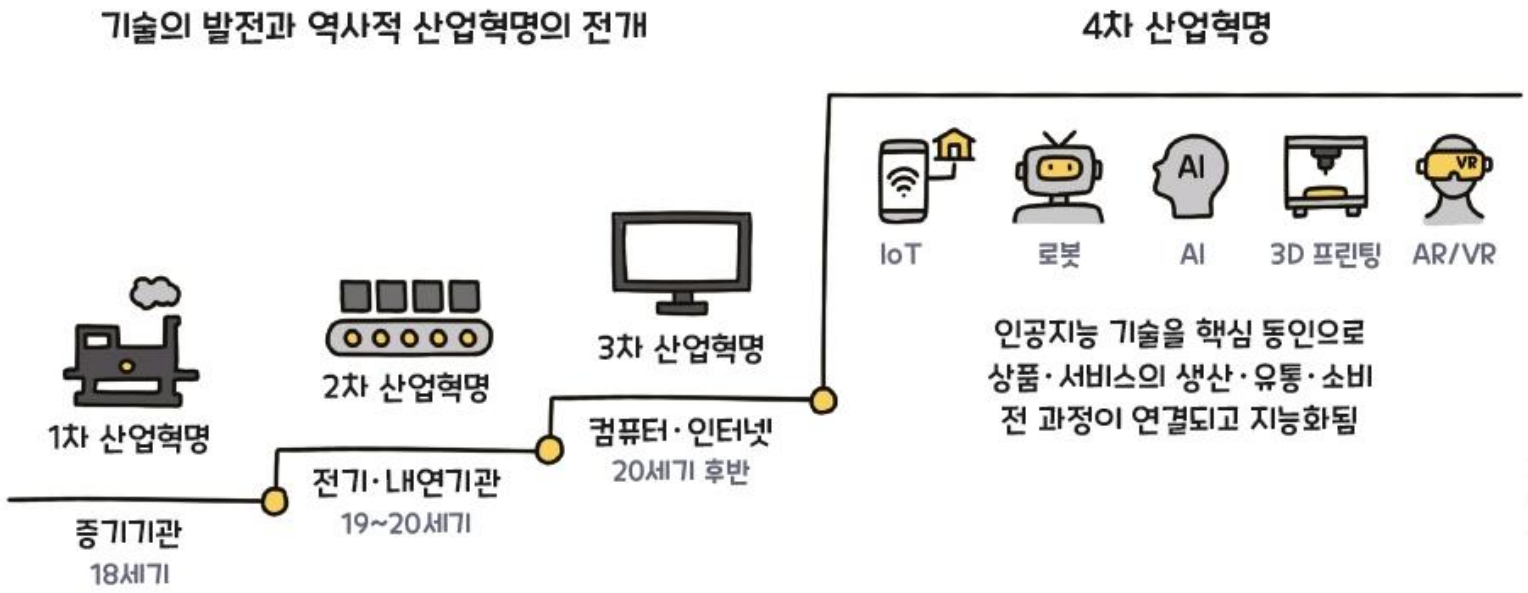
Reliability of AI's

❖ In a report to discuss ethical and social issues of artificial intelligence and autonomous systems, IEEE, described the changes that artificial intelligence will bring as follows



Reliability of AI's

❖ AI technology is the core of the 4th industrial revolution



- Is the result of AI with enough destructive power to change human life reliable?

Reliability of AI's

- Artificial Intelligence’s incomplete judgment
  - Accidents occur due to incomplete judgment of artificial intelligence
    - (예) 스코틀랜드 축구 경기에서 인공지능 카메라가 민머리인 심판을 축구공으로 오인하여 축구공이 아닌 심판만 계속 쫓아다닌 사건

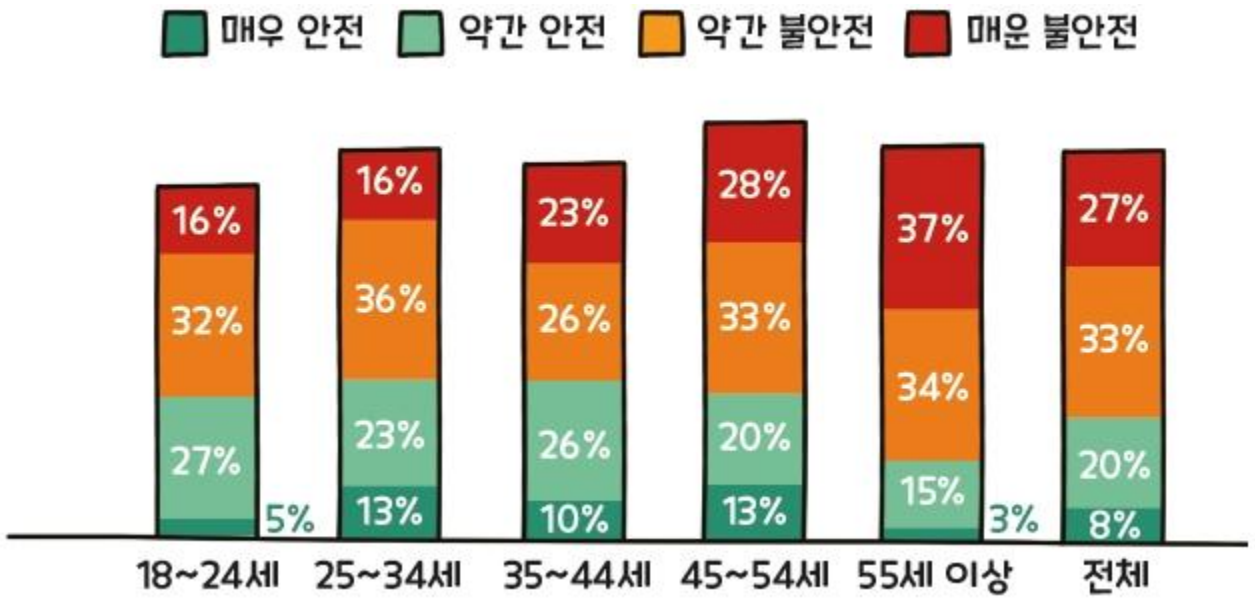
Reliability of AI's

- Artificial Intelligence’s incomplete judgment
  - Autonomous driving car
    - An accident in which a cyclist was hit and killed by Uber's self-driving car
    - Tesla's self-driving car was off the road, killing the driver

Reliability of AI's

❖ Artificial Intelligence’s incomplete judgment

- Autonomous driving car
  - Recently, as self-driving cars have been actively piloted, the number of drivers of rear-end and death accidents has begun to increase

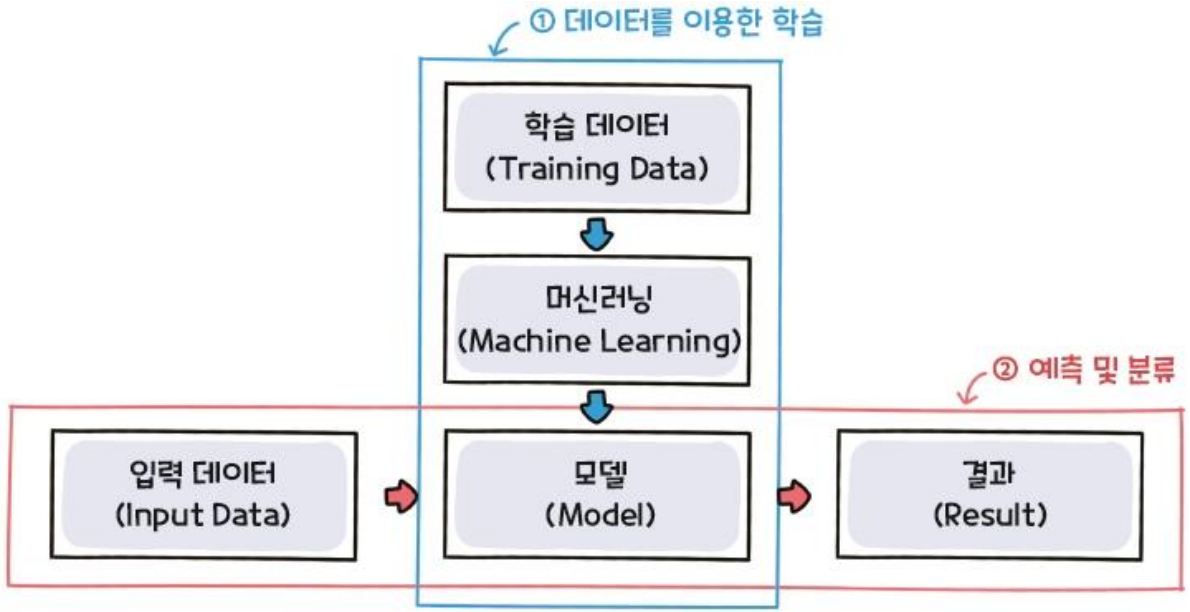


Reliability of AI's

- Artificial Intelligence’s incomplete judgment
  - Medical AI - IBM Watson
    - Indian hospital in Manipal determines rectal cancer at 85% matching rate
    - Watson's lung cancer diagnosis rate is only 17.8 percent
    - In breast cancer cases, non-reality matches 80%, but metastasis matches only 45%
    - Conclusion: Watson's cancer treatment project + artificial intelligence platform for new drug development → discontinuation/reduction

Reliability of AI's

- AI reliability issues start with training methods
  - Learning using data: Creating a model by learning a large amount of data given data
  - Prediction and classification: After the model is created, another new data is injected into the model to derive insights

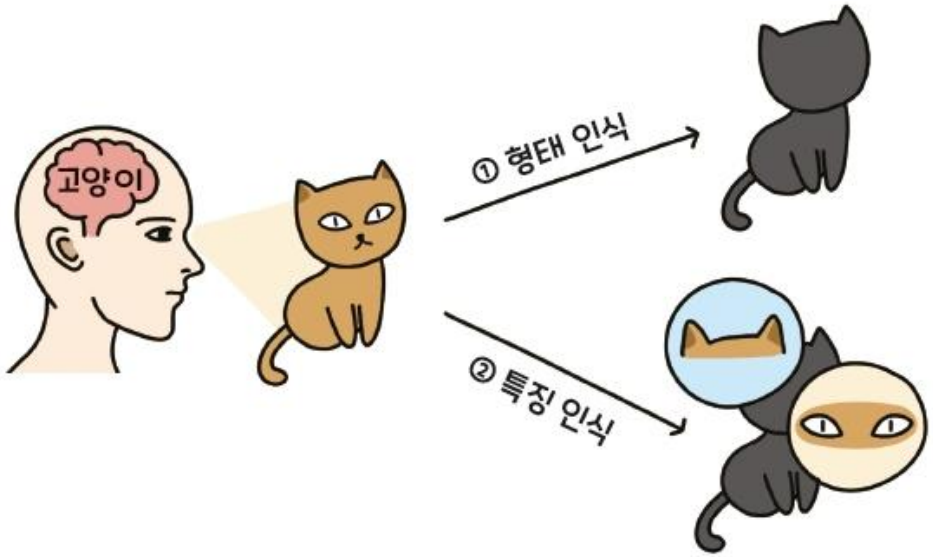


Reliability of AI's

- AI reliability issues
  - If the model is 100% accurate, you don't have to have doubts about the results
    - However, as AI makes judgments directly related to human life, doubts about whether AI can be fully trusted are amplified

Blackbox in AI

- Concepts
  - The way AI perceives things is not much different from humans because AI is also imitating human neural networks
  - How or on what basis the results were released by AI?
    - The inability to explain what process humans have gone through when they perceive an object as a 'cat'

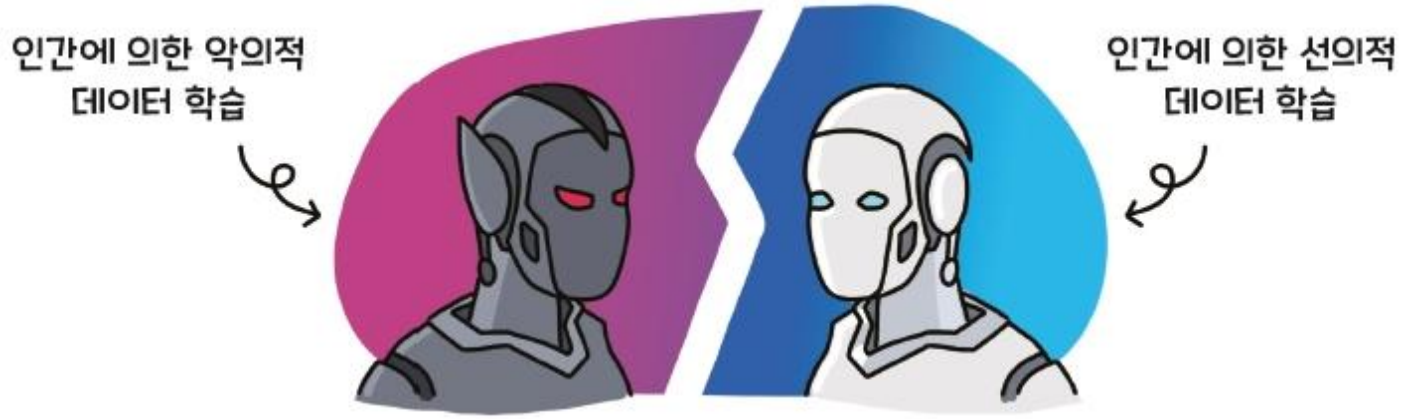


Blackbox in AI

- Concepts
  - The state of not being able to adequately explain the process or method of making a decision or decision made by an AI system
  - AI is completely complex that even experts cannot understand it, resulting in the so-called black box problem

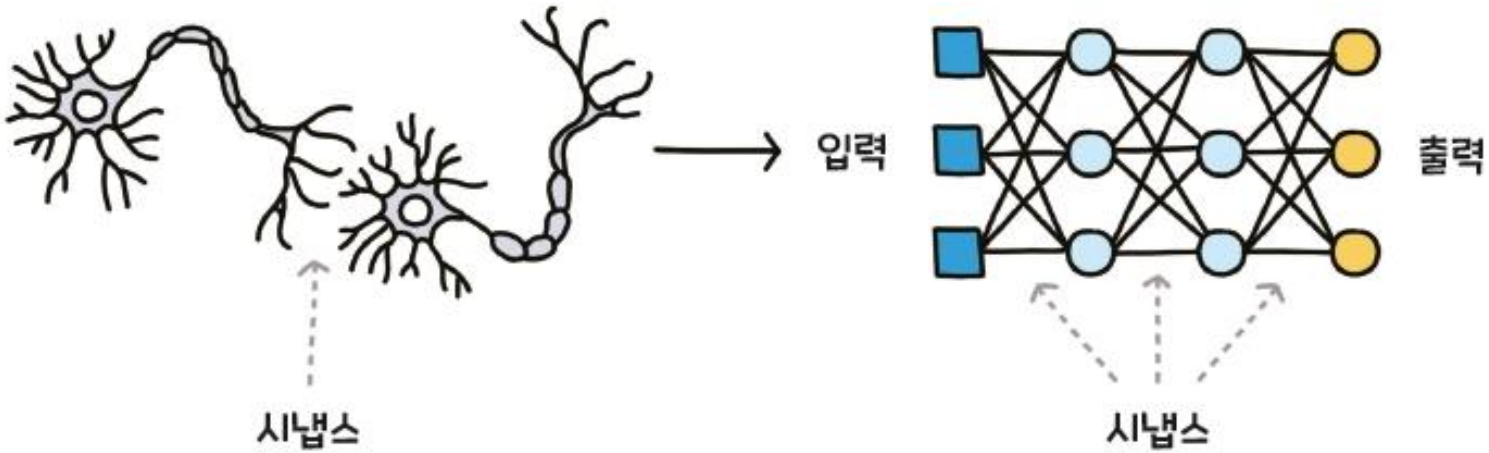
Blackbox in AI

- Reasons
  - Incompleteness of a training data
    - Since the learning data of AI comes from humans, it is objectivity ↓
    - Since bias is involved in the data raw material itself, it is inevitable that AI has a bias



Blackbox in AI

- Reasons
  - Uncertain data learning processes
    - Human inability to understand the learning process of AI
    - AI's neural network is a structure where millions of parameters interact in hundreds of layers of complexity
    - → virtually impossible for humans to recognize



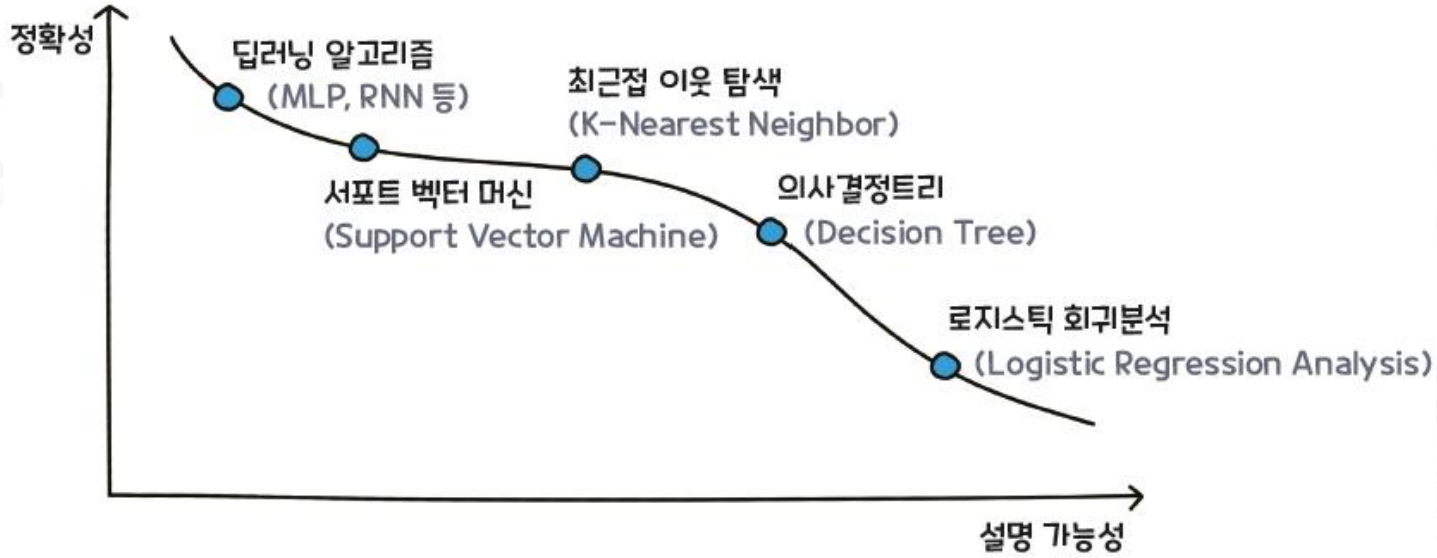
Blackbox in AI

❖ Reasons

- Uncertain data learning processes
  - 유럽연합은 개인정보보호 규정인 GDPR(General Data Protection Regulation)에서 유럽연합 시민은 프로파일링 등 자동화된 처리의 적용을 받지 않을 권리를 갖는다고 규정

Blackbox in AI

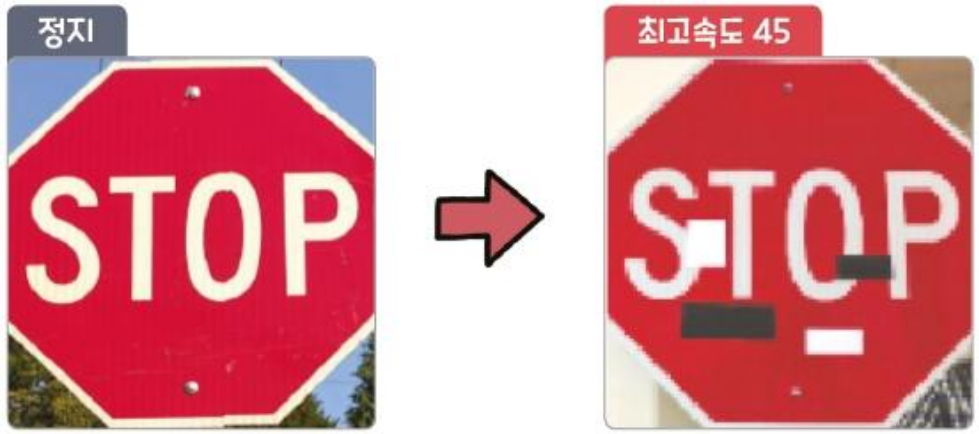
- Reasons
  - Uncertain data learning processes
    - If you can't understand the AI judgment process, you don't know if there's an error in the algorithm or what intention it was made: Trade-off
    - Realistically, you need a way to explain the process, even if it's not accurate



Explainable Artificial Intelligence (XAI)

❖ Background

- A case in which an autonomous vehicle that stops when looking at a "Stop" sign incorrectly judged that it was "Speed Limit 45" after seeing a "Stop" sign with noise data

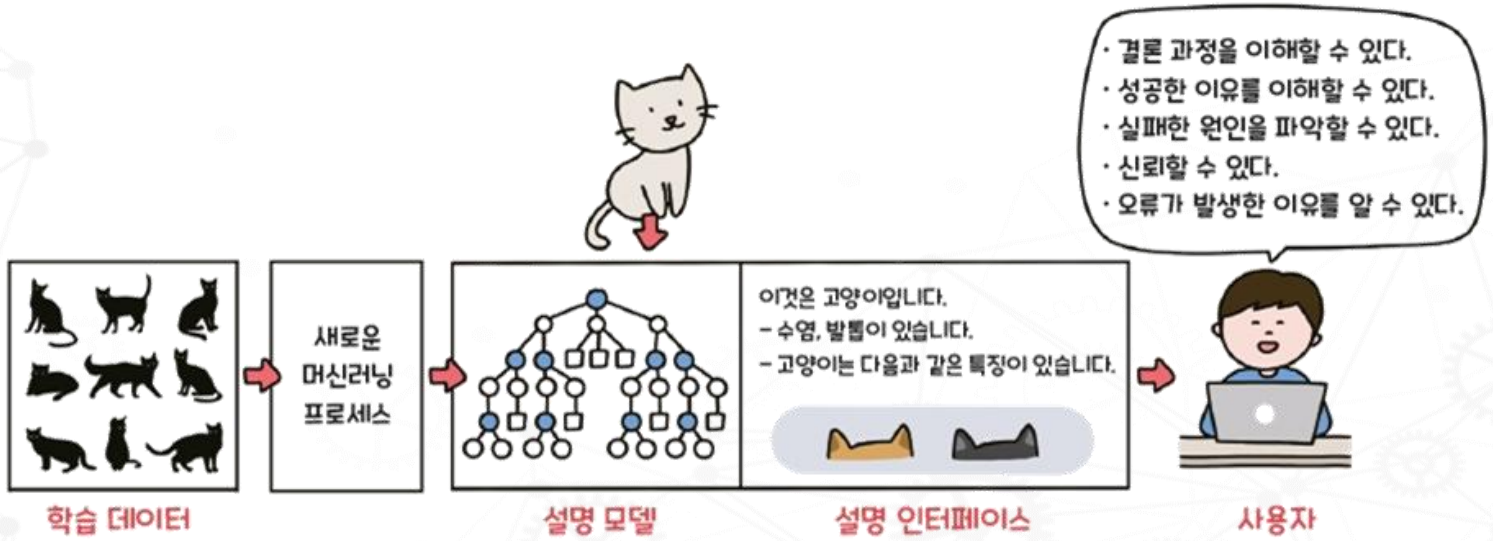


- To correct these errors, you need to know why and why AI is judging the results

Explainable Artificial Intelligence (XAI)

❖ Concepts

- Technology that explains the process by which results are generated so that users understand and correctly interpret the behavior and end-effects of AI systems
- AI that can explain the actions and judgments of AI in a form that humans understand

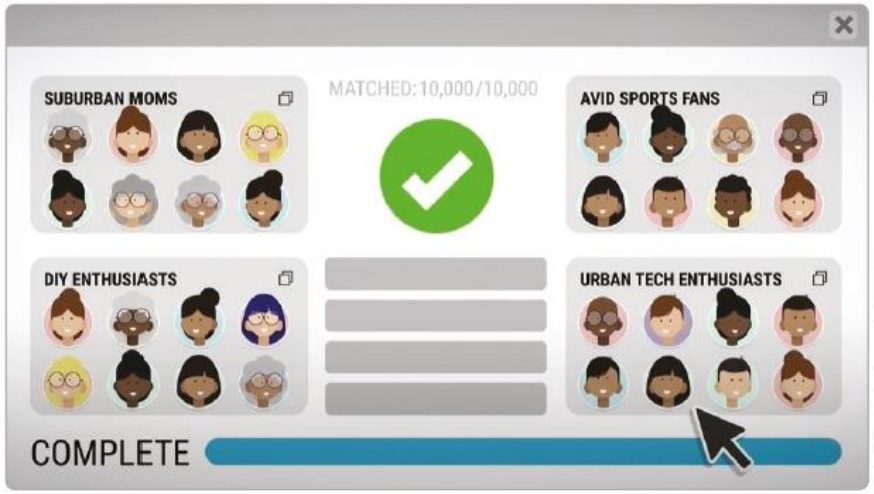


Explainable Artificial Intelligence (XAI)

- Concepts
  - Currently, AI injects learning data into the learning model, analyzes it, and delivers the results to the user, and in the learning area, the probability value is calculated and delivered to the user
    - Since the probability of object (cat) analysis is 93%, it is suggested to the user that it is a "cat", but in the process, humans do not know why AI has produced these results

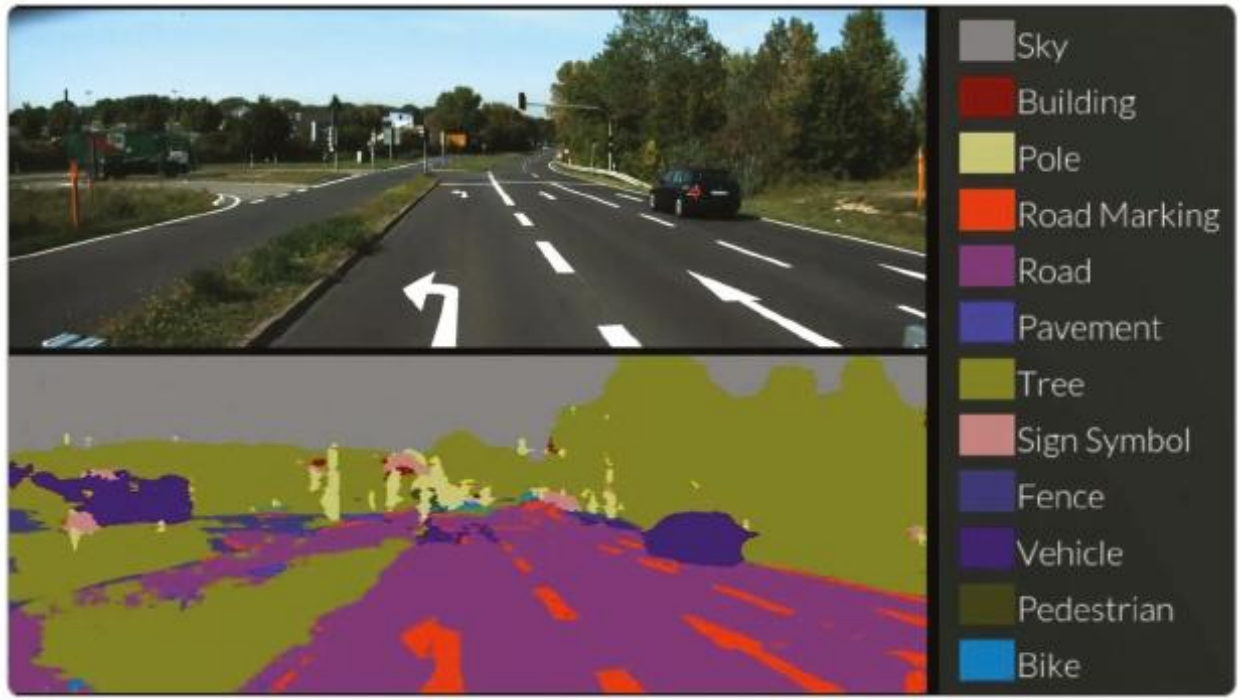
Explainable Artificial Intelligence (XAI)

- Applications
  - AI simMachines
    - Provides a visual indication of how accurate predictions can be made
    - Time series analysis explains changes over time and causes of changes at intervals
    - Example of dynamically classifying predictions by applying AI sim machines to marketing



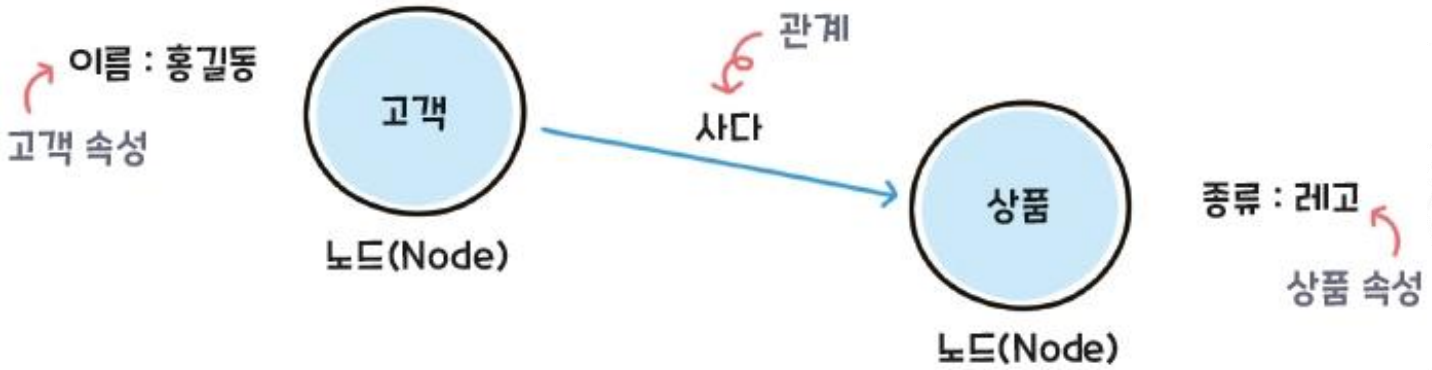
Explainable Artificial Intelligence (XAI)

- Applications
  - Autonomous driving car
    - Visualizing the way things are perceived



Explainable Artificial Intelligence (XAI)

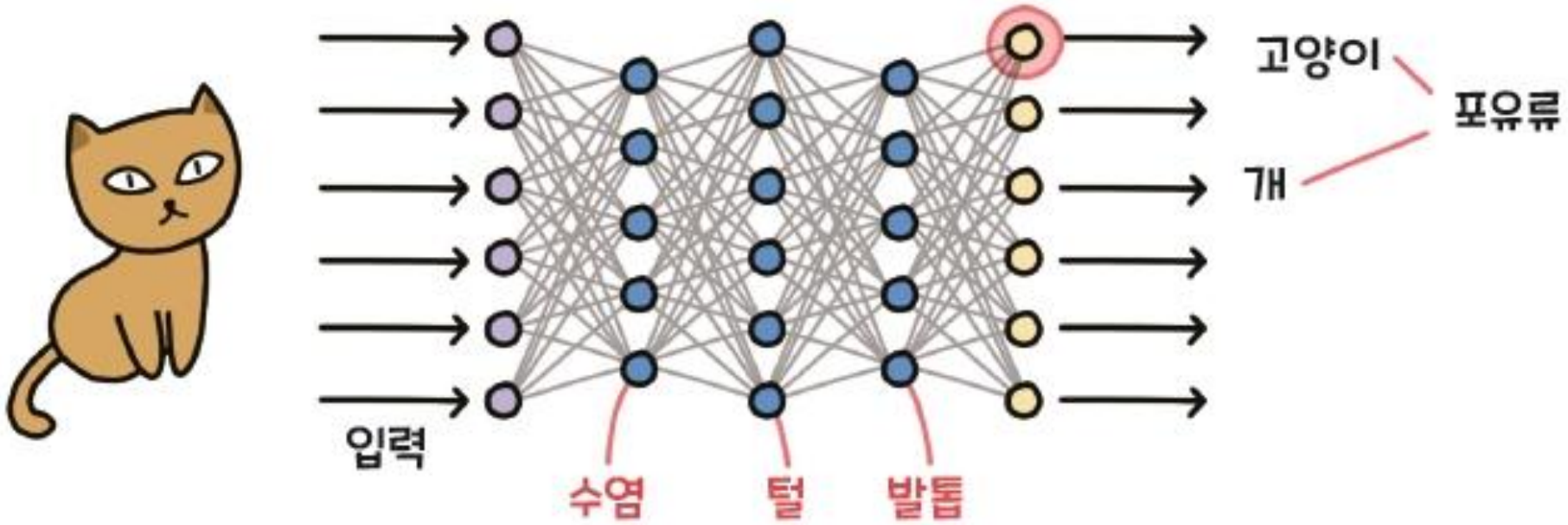
- How to visualize for XAI
  - The U.S. Advanced Research Planning Agency (DARPA) published a study that found that one of the interpretable models needed for explainable artificial intelligence (XAI) was a graph-based model
  - Graphs are data visualization representations using dots and lines, and explainable artificial intelligence can be implemented by visually expressing the learning process



Explainable Artificial Intelligence (XAI)

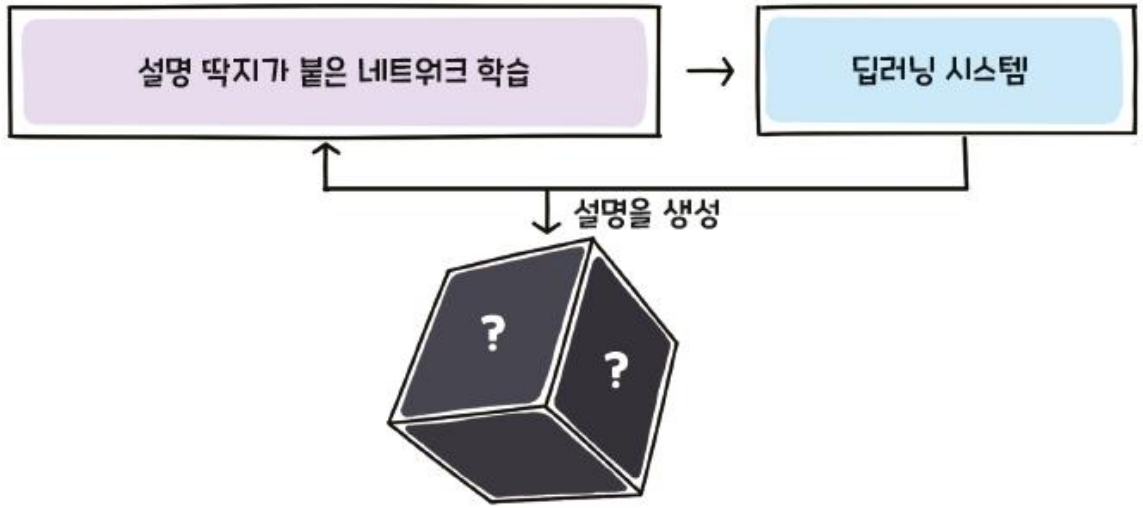
❖ How to visualize for XAI

- Attaching a description to a neural network node
  - (예) 인공신경망은 고양이의 수염, 털, 발톱과 같은 이미지의 특정 부분에 특정 노드를 지정하고 퍼즐 맞추기 게임처럼 모든 노드를 조합하여 대상 인식



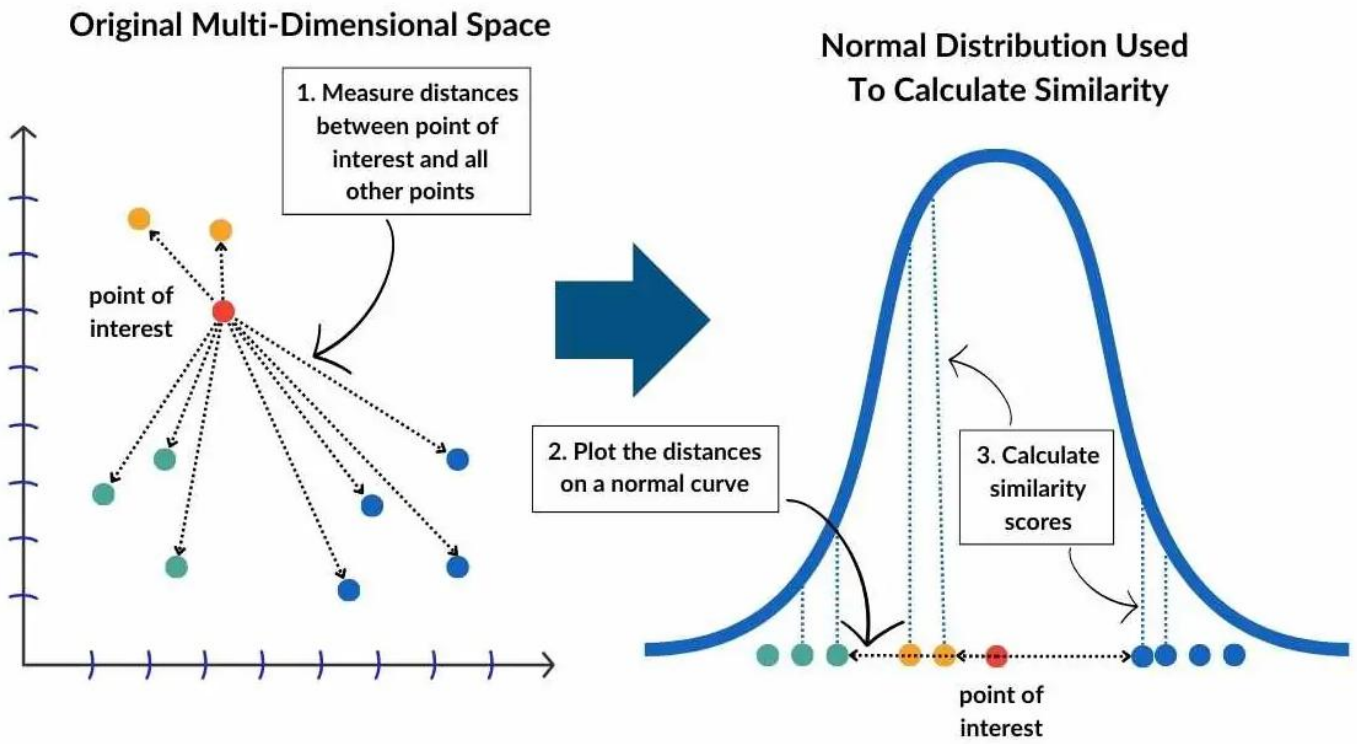
Explainable Artificial Intelligence (XAI)

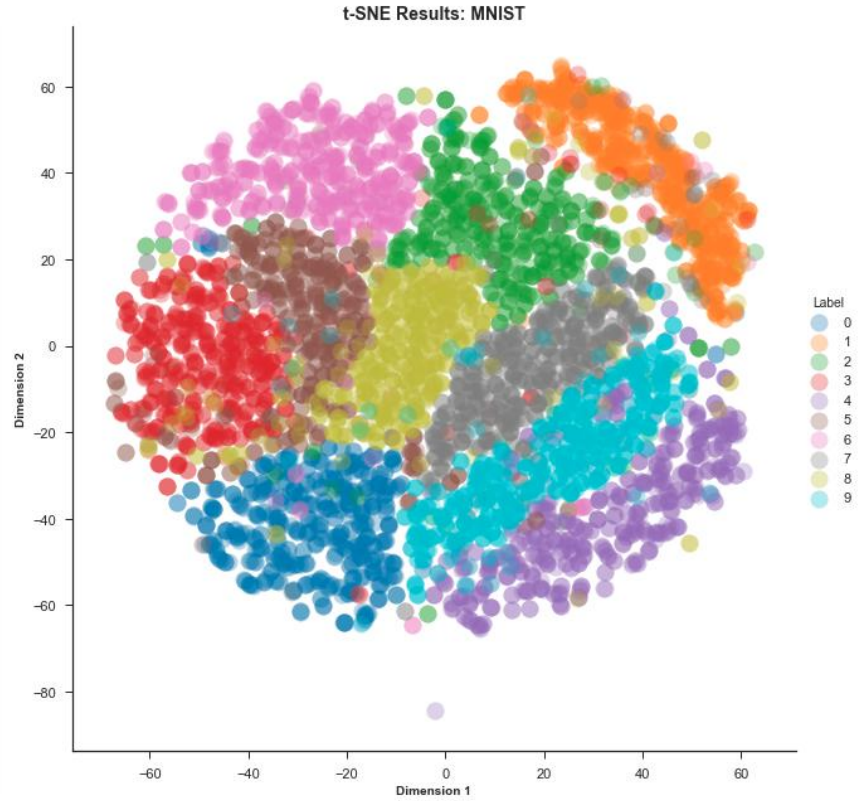
- How to visualize for XAI
  - To infer a model
    - Learn network labeled description
    - Train the deep learning system to explain how the system reached its final conclusion



Explainable Artificial Intelligence (XAI)

- How to visualize for XAI
  - T-SNE (t-distributed stochastic neighbor embedding)





Explainable Artificial Intelligence (XAI)

- How to visualize for XAI
  - GradCAM (Gradient-weighted Class Activation Mapping)

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

Abstract We propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent and explainable.

Our approach – Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (*e.g.* VGG), (2) CNNs used for structured outputs (*e.g.* captioning), (3) CNNs used in tasks with multi-modal inputs (*e.g.* visual question answering) or reinforcement learning, all *without architectural changes or re-training*. We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures.

In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are robust to adversarial perturbations, (d) are more faithful to the underlying model, and (e) help achieve model generalization by identifying dataset bias.

For image captioning and VQA, our visualizations show that even non-attention based models learn to localize discriminative regions of input image.

We devise a way to identify important neurons through Grad-CAM and combine it with neuron names [4] to provide textual explanations for model decisions. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a 'stronger' deep network from a 'weaker' one even when both make identical predictions. Our code is available at https://github.com/ramprs/grad-cam/, along with a demo on CloudCV [2][1], and a video at youtu.be/COjUB9Izk6E.

Ramprasaath R. Selvaraju
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: ramprs@gatech.edu

Michael Cogswell
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: cogswell@gatech.edu

Abhishek Das
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: abhshkdz@gatech.edu

Ramakrishna Vedantam
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: vrama@gatech.edu

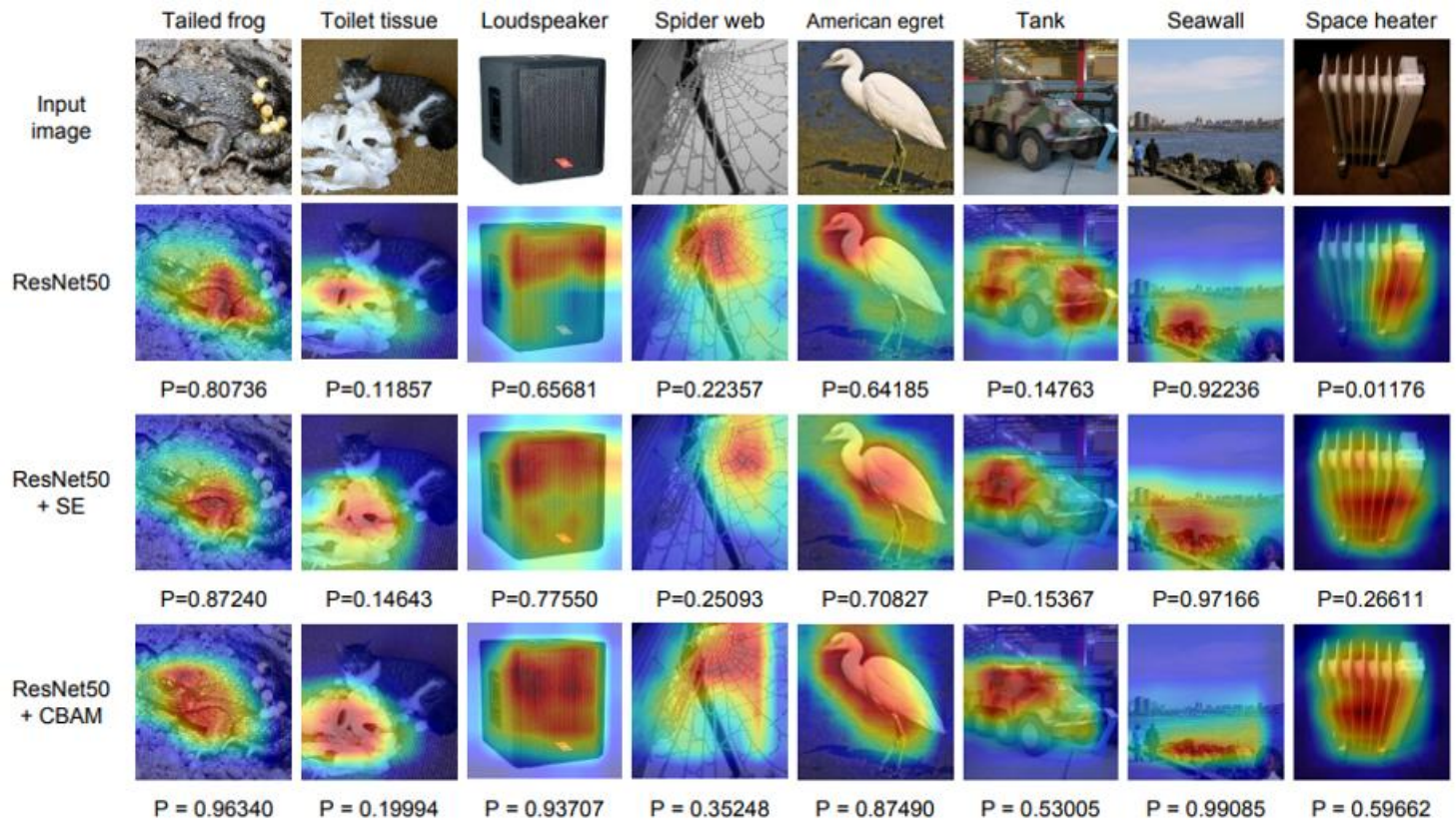
Devi Parikh
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA
E-mail: parikh@gatech.edu

Dhruv Batra
Georgia Institute of Technology, Atlanta, GA, USA
Facebook AI Research, Menlo Park, CA, USA
E-mail: dbatra@gatech.edu

1 Introduction

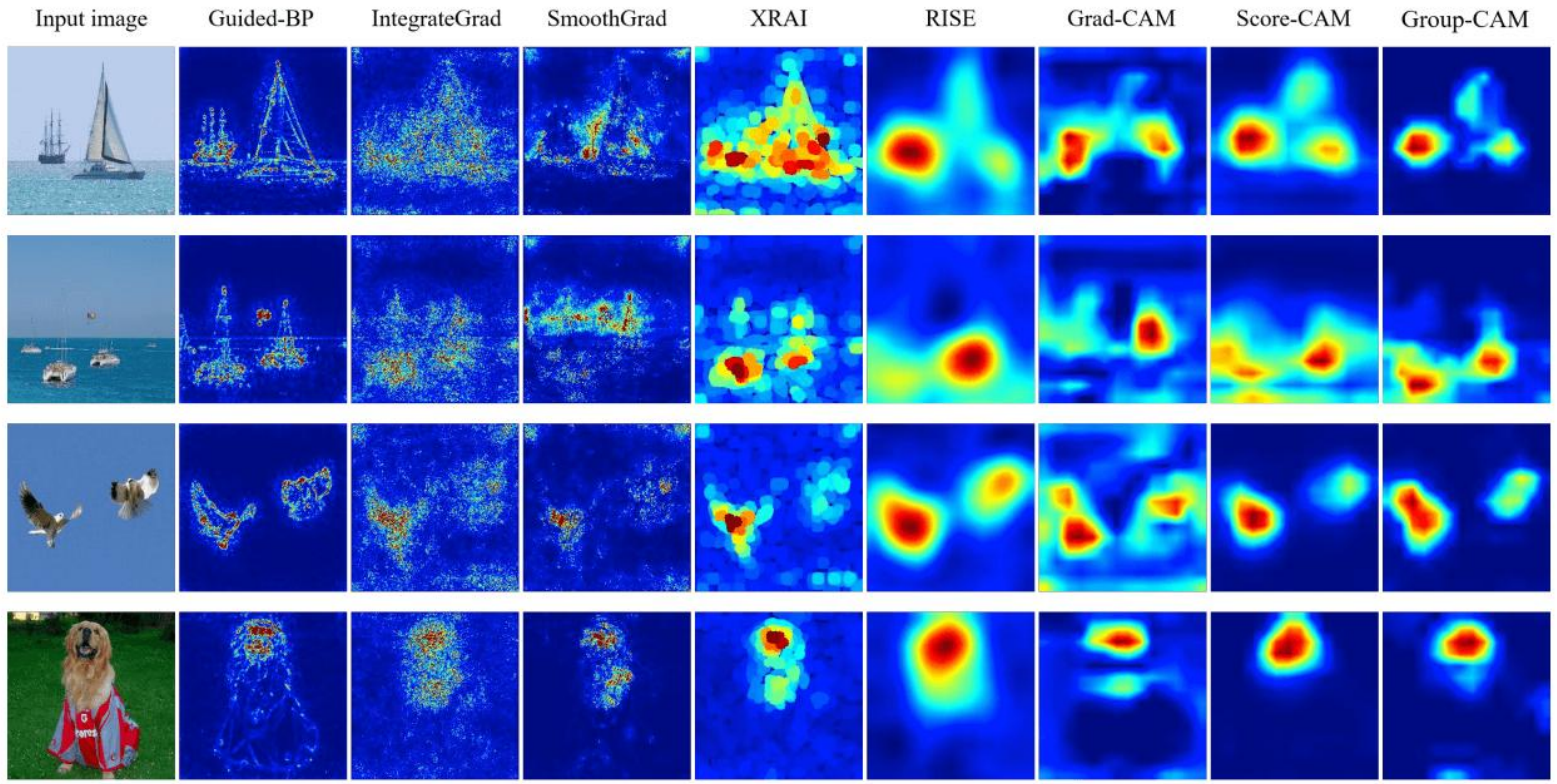
Deep neural models based on Convolutional Neural Networks (CNNs) have enabled unprecedented breakthroughs in a variety of computer vision tasks, from image classification [33,24], object detection [21], semantic segmentation [37] to image captioning [55,7,18,29], visual question answering [3,20,42,46] and more recently, visual dialog [11,13,12] and embodied question answering [10,23]. While

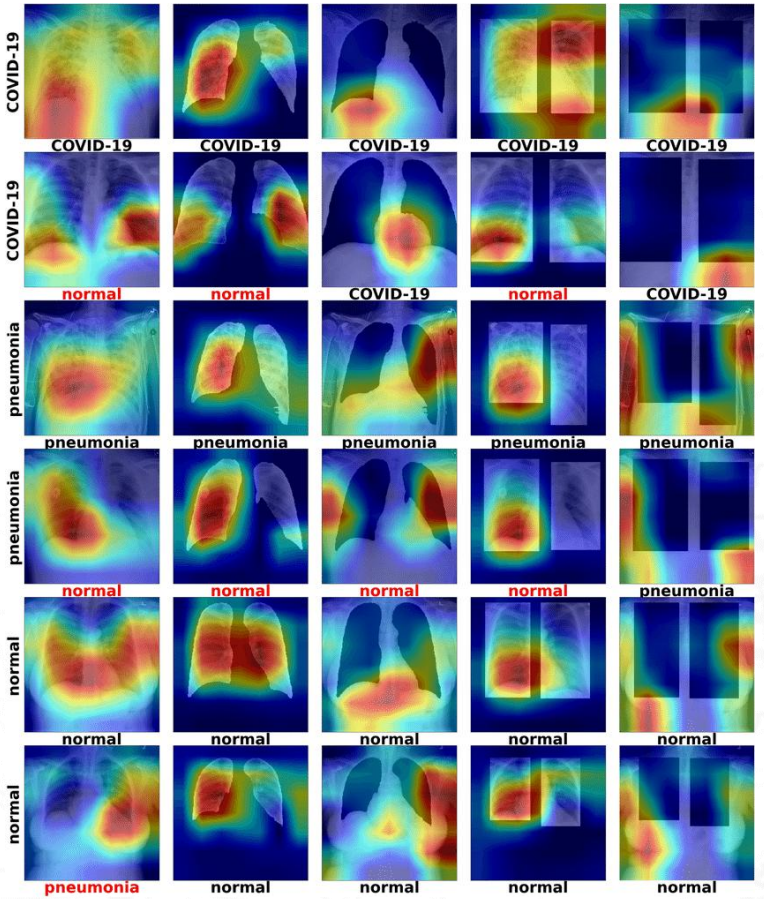
[1] http://gradcam.cloudcv.org



Explainable Artificial Intelligence (XAI)

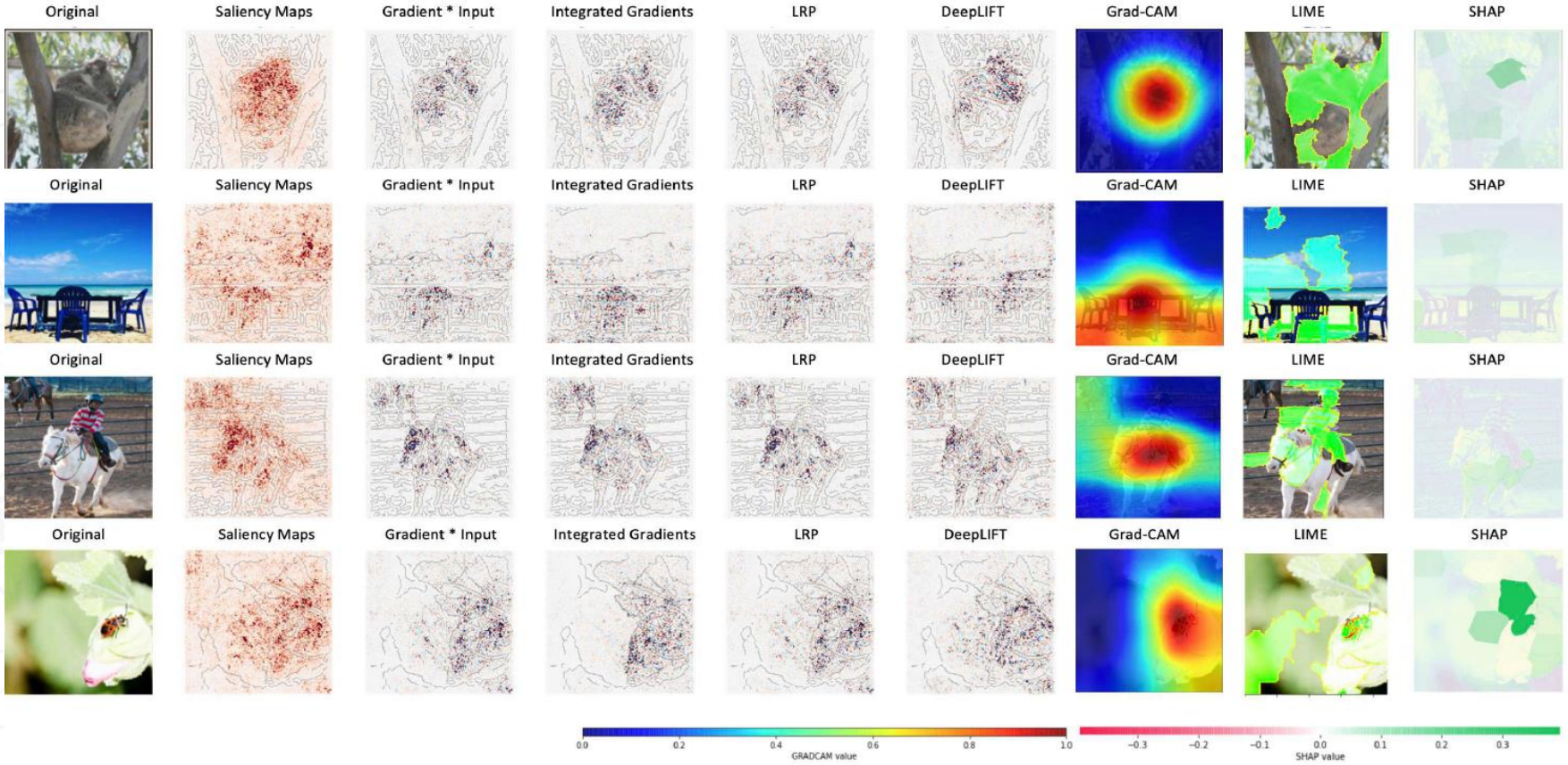
- How to visualize for XAI
  - GradCAM (Gradient-weighted Class Activation Mapping)

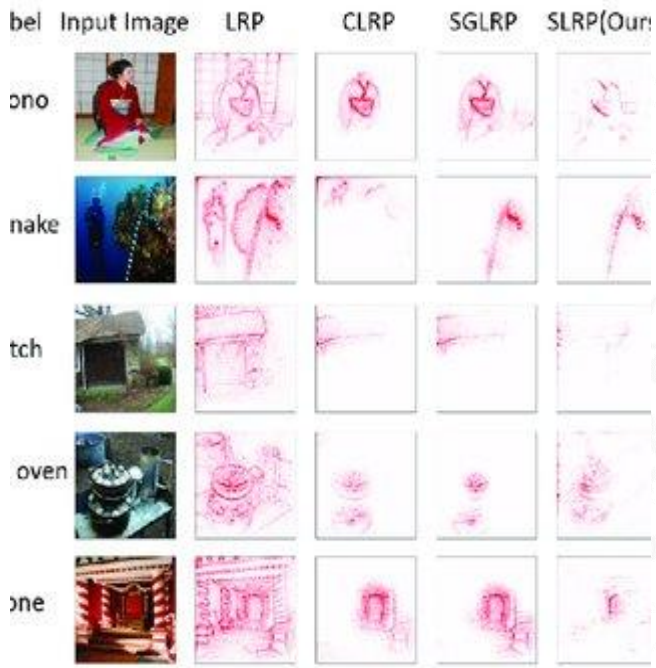




Explainable Artificial Intelligence (XAI)

- How to visualize for XAI
  - LRP (Gradient-weighted Class Activation Mapping)





Explainable Artificial Intelligence (XAI)

❖ AI capability assessment elements provided by DARPA

항목	평가지표
사용자 만족도	• 설명이 얼마나 명확한가? • 설명이 얼마나 유용한가?
설명모델 수준	• 개별 의사결정 이해도 • 전체 모델에 대한 이해도 • 장단점 평가 • '미래 해동' 예측 • '개입 방법' 예측
업무수행 향상도	• 설명이 사용자 의사결정, 업무수행 능력을 향상시켰는가? • 사용자의 이해도 평가를 위한 실험적 업무
신뢰성 평가	• 미래에도 사용할 만큼 신뢰하는가?
오류 수정 수준(가점)	• 인식 오류 수준 • 인식 오류 수정을 위한 지속적인 훈련

Ethics for AI

Singularity of AI

- Concepts
  - A time when AI transcends human intelligence, when humans cannot keep up with the development of technology

Singularity of AI

- Concepts
  - In the era of reaching singularity, computers will design and produce AI that is more advanced than themselves,
    - and that will be able to design and produce another AI that has better intelligence than themselves
  - The reason why we need to prepare and prepare for singularities is related to control over AI

Singularity of AI

❖ Issues of singularity

- The controversy over AI singularities intensifies as the development of AI begins in earnest
- Scholars are divided into anti-communists and pro-communists

Singularity of AI

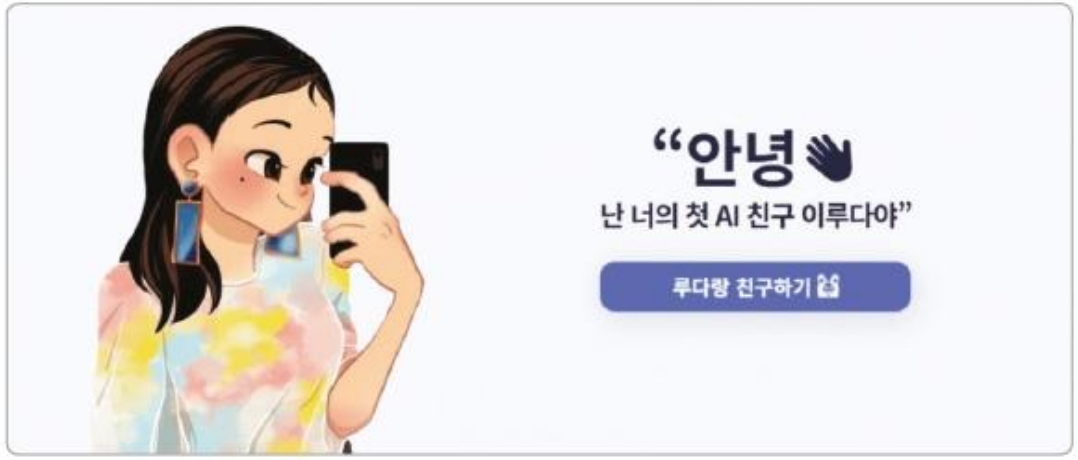
❖ Issues of singularity

하나 더 알기 **인공지능의 통제 : 킬 스위치**

- **킬 스위치(Kill Switch) :**
  인공지능이 이상반응을 보일 때 외부에서 강제로 종료시키는 것
- 구글은 인공지능이 인간의 의도나 통제를 벗어나 해를 끼칠 수 있는 '반역'이 예상되는 상황일 때 즉각 안전 정지시켜 위험을 차단하는 대비책을 마련하기 위한 연구를 계속하고 있음

Need for AI Ethics

- Examples
  - Case 1: Chatbot ‘이루다’
    - Some users exploit the learning ability of ‘이루다' to inject inappropriate words
    - The situation in which ‘이루다' makes hate speech without adding or subtracting it
      - Out of service within 20 days of service start



Need for AI Ethics

❖ Examples

- Case 2: ProPublica predicts recidivism rate
  - AI deduces the recidivism rate of blacks higher than that of whites
- Case 3: Amazon recruitment AI
  - As the issue of discrimination against women in recruitment programs using AI emerged, the program was discarded on its own

(a) 인공지능 판결

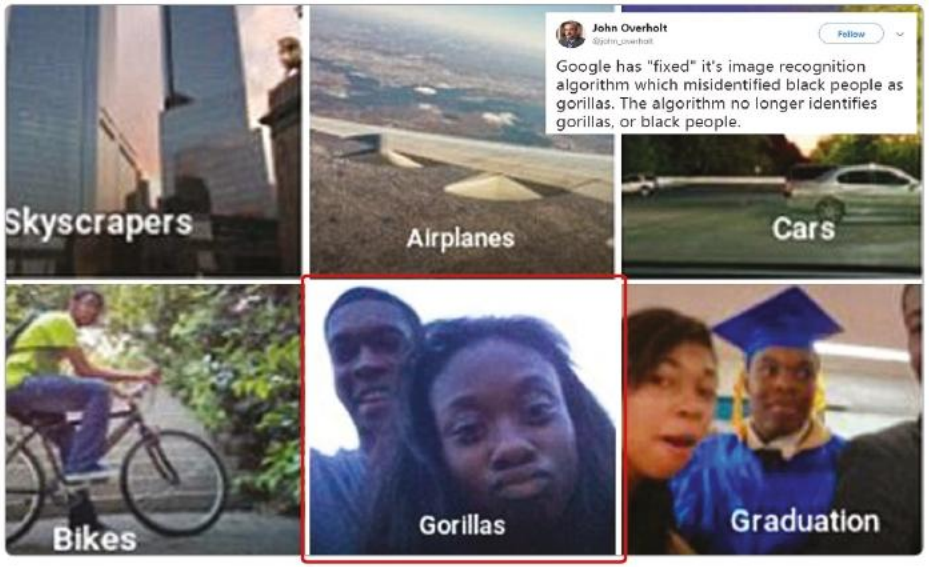
(b) 인공지능 면접

Need for AI Ethics

- Necessary
  - Satya Nadella, CEO of Microsoft, brings up the topic of AI ethics

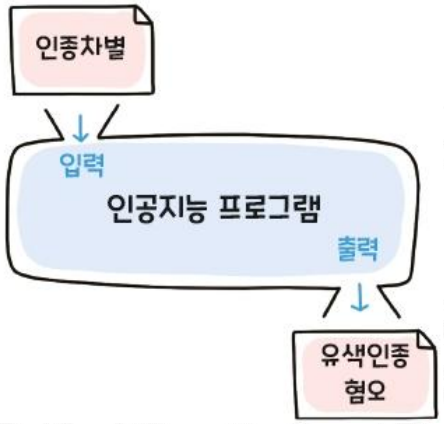
Need for AI Ethics

- Necessary
  - Google’s image recognition
    - Recognizing black women as gorillas during image recognition
  - WeChat translation course
    - Use the word Negro



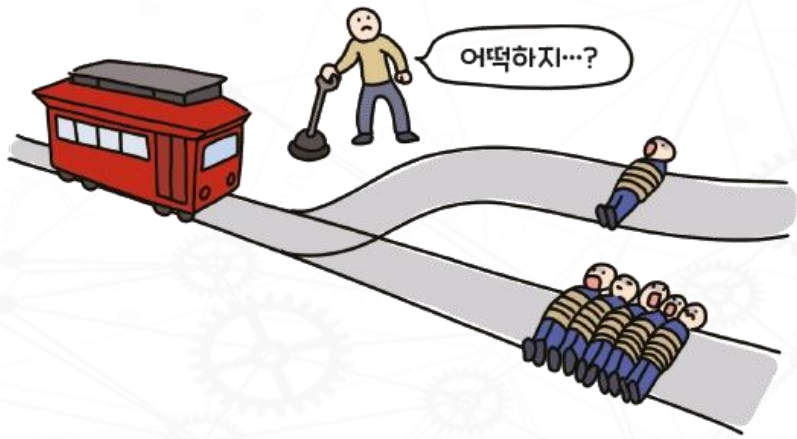
Need for AI Ethics

- Necessary
  - MIT Media Lab presentation
    - The probability that artificial intelligence will cause errors in the process of recognizing white male faces is 1%
    - Black women have a 35% chance of error
  - AI algorithms have different results depending on what data you insert
    - What data to inject into AI is up to humans



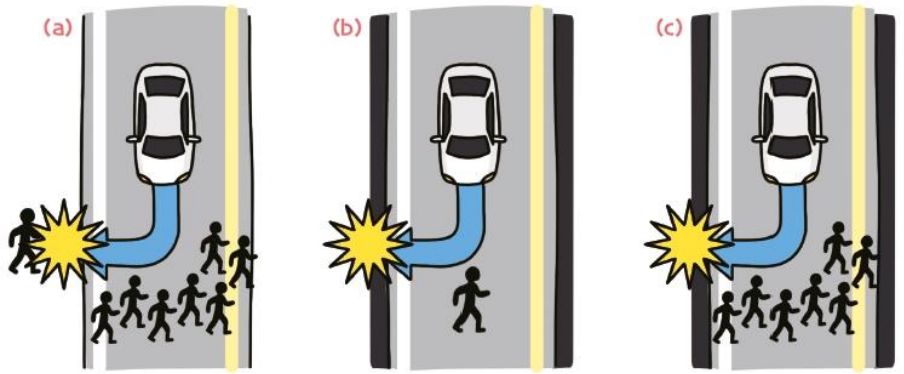
Need for AI Ethics

- Trolley dilemma
  - 윤리학 분야의 사고실험 중 하나
    - 다섯 사람을 구하기 위해 한 사람을 죽이는 것이 도덕적으로 허용 가능한지에 대한 질문
    - 트롤리 전차가 철길 위의 5명의 인부들을 향해 빠른 속도로 돌진
    - 당신 옆에 트롤리의 방향을 바꿀 수 있는 레일 변환기가 있음
      - 방향을 왼쪽으로 바꾼다면 왼쪽 철로에서 일하는 1명의 인부 사망
      - 방향을 바꾸지 않는다면 5명의 인부들 사망



Need for AI Ethics

- Trolley dilemma
  - For the autonomous driving car



  - (a) : 여러 사람이 희생되는 것보다는 한 사람이 희생되는 것이 올바른 선택 같아 보이기 때문에 인공지능의 선택은 그리 어렵지 않을 것
  - (b) : 자율주행차는 그냥 보행자를 치고 지나가야 할까, 아니면 운전자가 다치게끔 방향을 꺾어야 할까?
  - (c) : 여러 사람의 목숨과 운전자의 목숨 중 자율주행차는 어느 쪽에 더 비중을 두고 판단을 내려야 할까?

Need for AI Ethics

- Trolley dilemma
  - The most basic ethical issues to face in the era of self-driving car
  - Need to think about the ethics of algorithms before it's too late

AI Ethics

- Asilomar AI Principles
  - 23 rules for developers to observe on the purpose, ethics, value, etc. of AI advances
    - This principle consists of three parts: research-related issues, ethics and values, and long-term issues

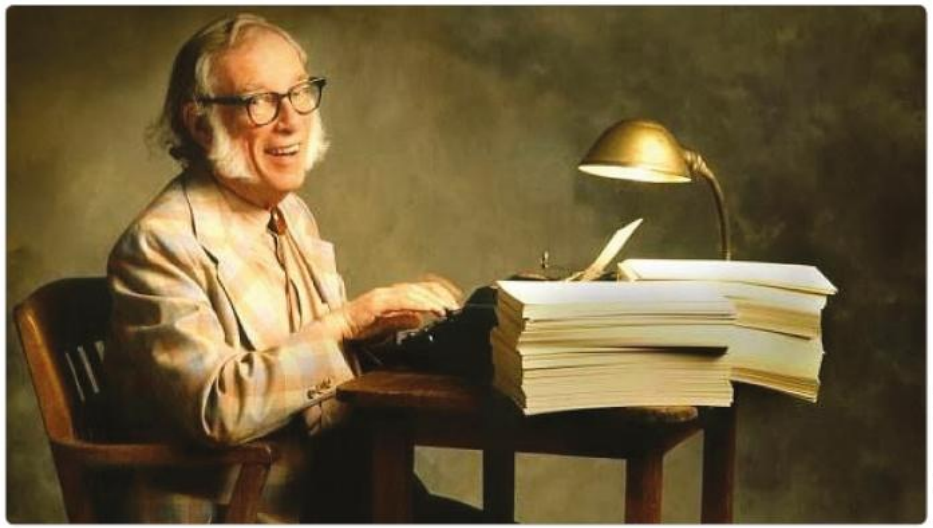
AI Ethics

❖ Asilomar AI Principles

1) **연구 관련 쟁점**
   - 연구 목표, 연구비 지원, 과학정책 연계, 연구 문화, 경쟁 회피 등
2) **윤리와 가치**
   - 안전, 실패의 투명성, 사법적 투명성, 책임성, 가치 일치, 인간의 가치, 개인정보보호, 자유와 프라이버시, 이익의 공유, 번영의 공유, 인간 통제, 사회전복 방지, 인공지능 무기 경쟁 지양 등
3) **장기적 이슈**
   - 역량 경고, 중요성, 위험성, 자기개선 순환, 공동의 선 등

AI Ethics

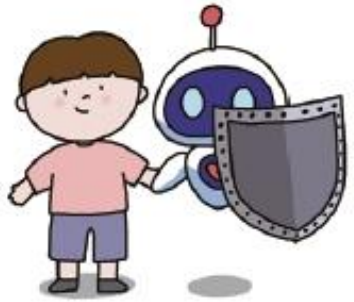
- Three laws of robotics
  - Three principles that robots must follow
    - Issac Asimov, the author who suggested the three principles of robots, believes that robots will not be a threat to humans if they follow these principles well



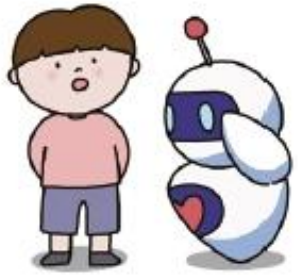
AI Ethics

❖ Three laws of robotics

- **제1원칙 :** 로봇은 인간에게 해를 입혀서는 안 되고, 위험에 처한 인간을 방치해서도 안 된다.
- **제2원칙 :** 제1 원칙을 어기지 않는 한, 로봇은 인간의 명령에 복종해야 한다.
- **제3원칙 :** 제1 원칙과 제2원칙을 어기지 않는 한, 로봇은 로봇 자신을 지켜야 한다.



제1원칙 로봇은 인간에게 해를 입혀서는 안 되고, 위험에 처한 인간을 방치해서도 안 된다.



제2원칙 제1원칙을 어기지 않는 한, 로봇은 인간의 명령에 복종해야 한다.

제3원칙 제1원칙과 제2원칙을 어기지 않는 한, 로봇은 로봇 자신을 지켜야 한다.

AI Ethics

- Three laws of robotics
  - Since then, Isaac Asimov has proposed additional robot ‘0 principle’ in his short story “Robots and Empire”
  - 0 principle for robots : 로봇은 인류에게 해를 가할 만한 명령을 받거나 행동을 하지 않음으로써 '인류'에게 해가 가해지는 것을 방치해서도 안 된다 (Extend to 1st principle)