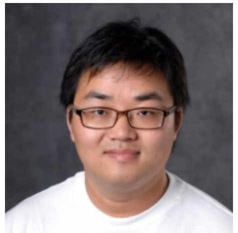
Trustworthy AI:
A Computational Perspective



Haochen Liu[1], Yiqi Wang[1], Wenqi Fan[2], Xiaorui Liu[1], Yaxin Li[1] and Jiliang Tang[1]

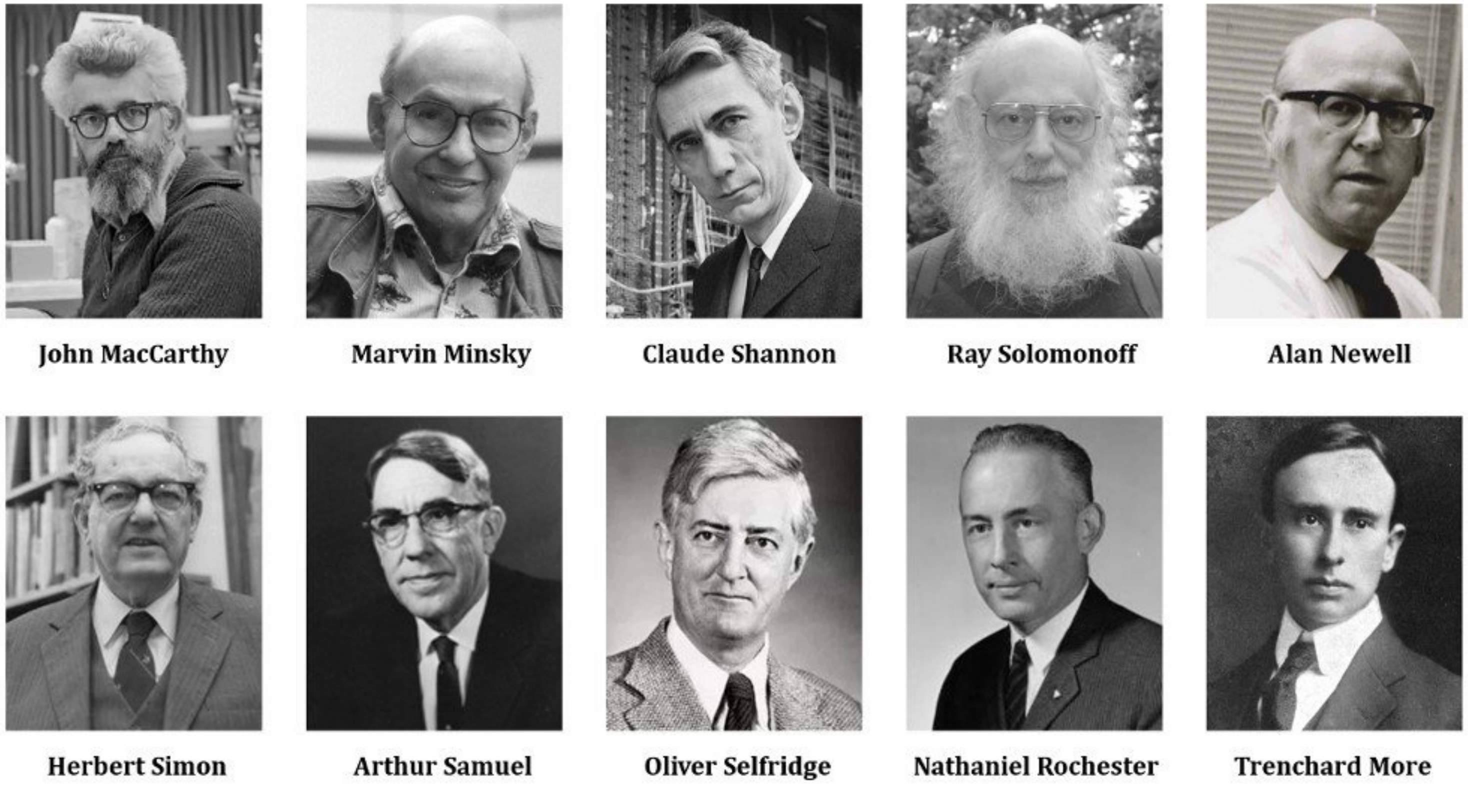
[1]Michigan State University

[2]The Hong Kong Polytechnic University

Tutorial website: [Trustworthy AI: A Computational Perspective]

Artificial Intelligence (AI)

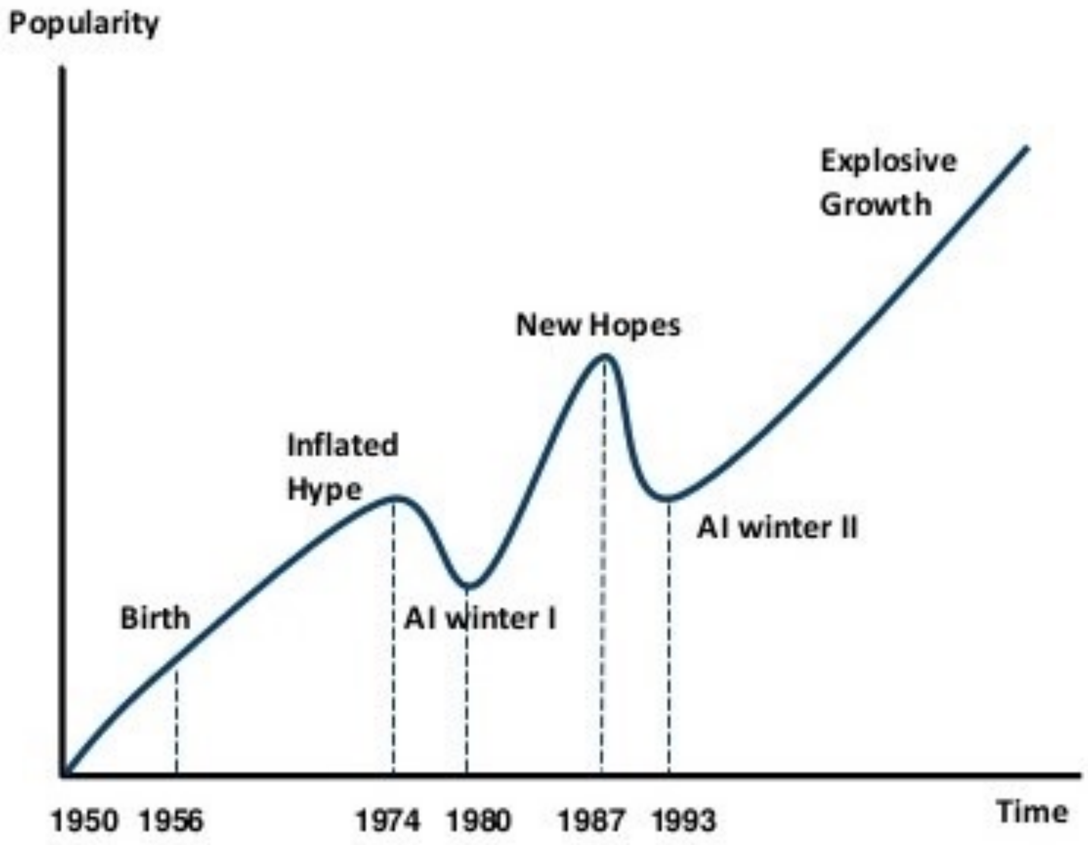
1956 Dartmouth Conference:
The Founding Fathers of AI



A program or a system which is able to cope with a real-world problem with humanlike reasoning capability.

AI Summers and Winters

AI HAS A LONG HISTORY OF BEING "THE NEXT BIG THING"...



Timeline of AI Development

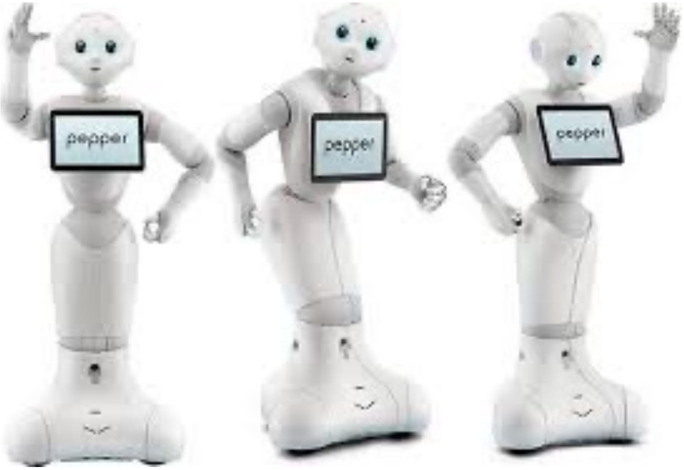
- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM's Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world's champions

https://www.actuaries.digital/2018/09/05/history-of-ai-winters/

AI is Everywhere

Business

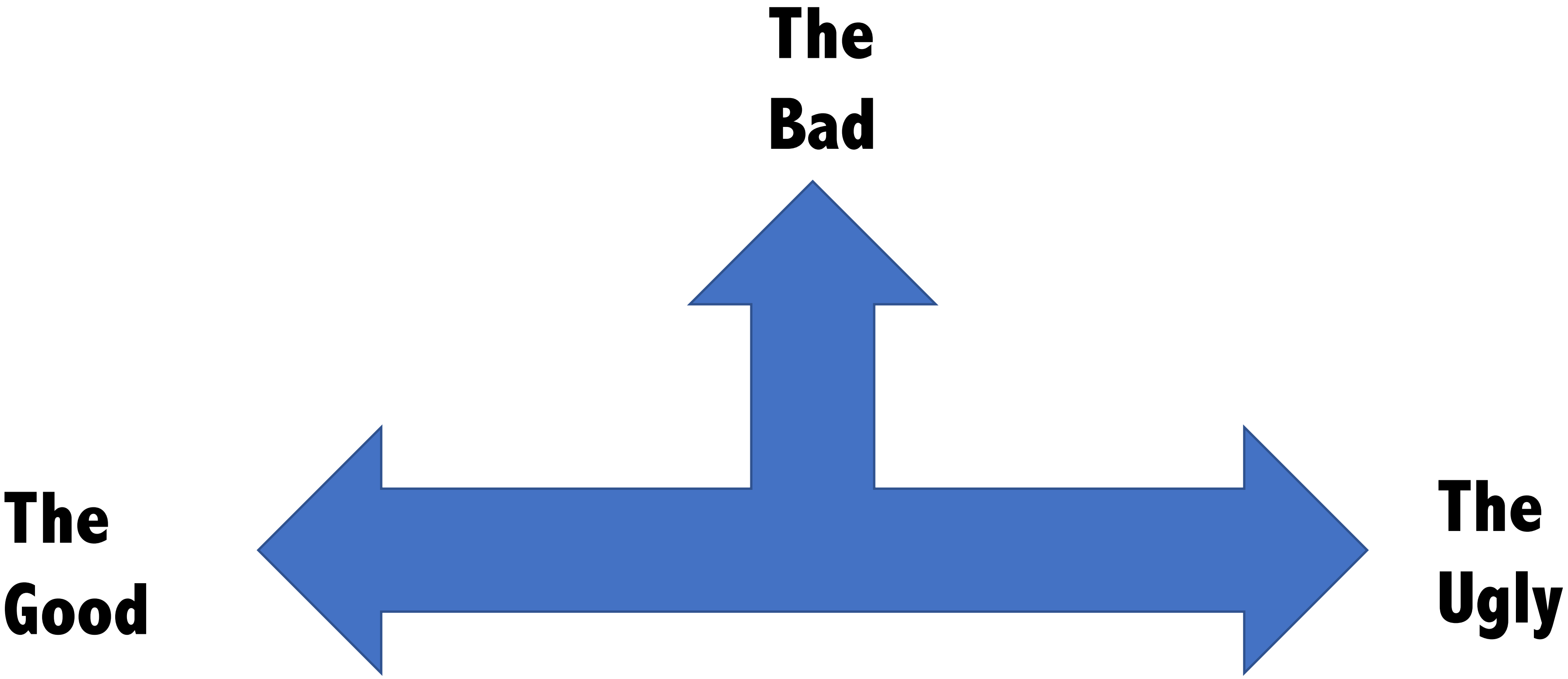
Healthcare



Robotics

Education

The Good, The Bad, and The Ugly

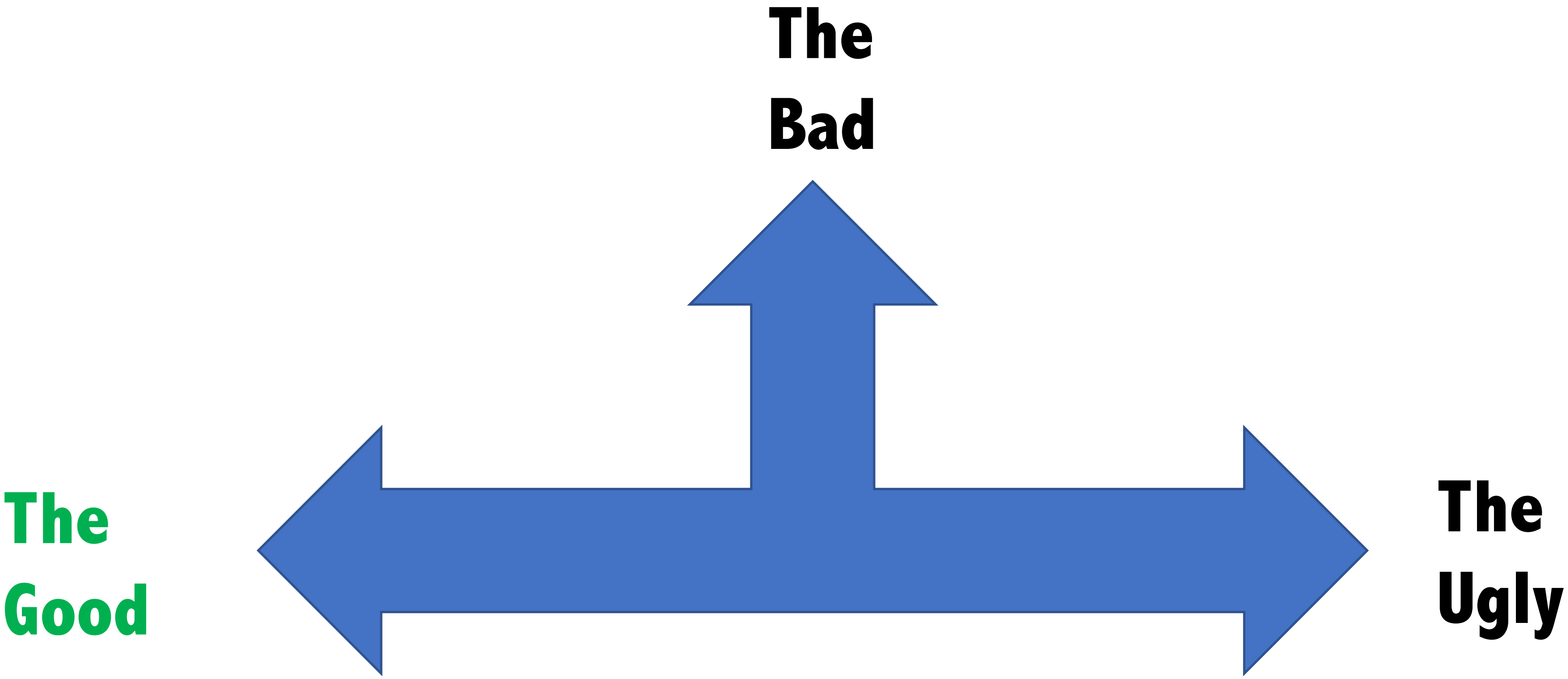


The Bad

The Good

The Ugly

The Good, The Bad, and The Ugly



The Bad

The Good

The Ugly

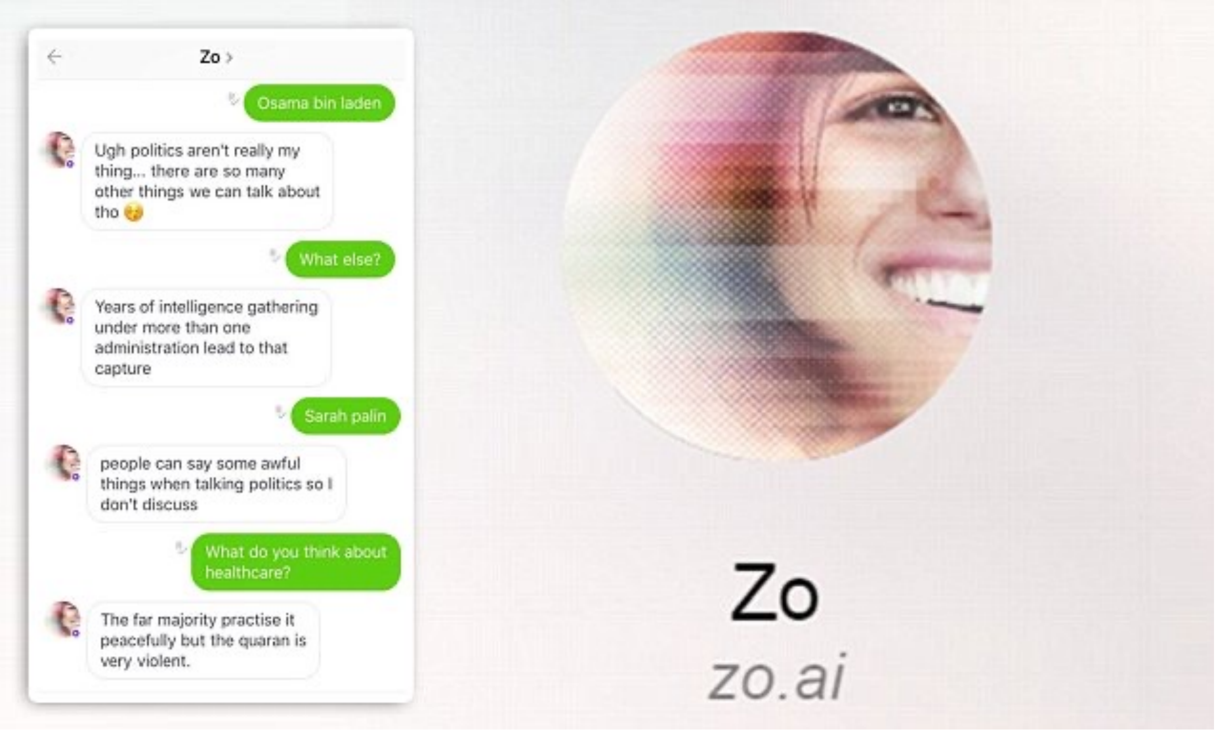
Face Recognition

Criminal Identification

Face ID

Conversational AI

Voice Assistant



Osama bin laden

Ugh politics aren't really my thing... there are so many other things we can talk about tho 😔

What else?

Years of intelligence gathering under more than one administration lead to that capture

Sarah palin

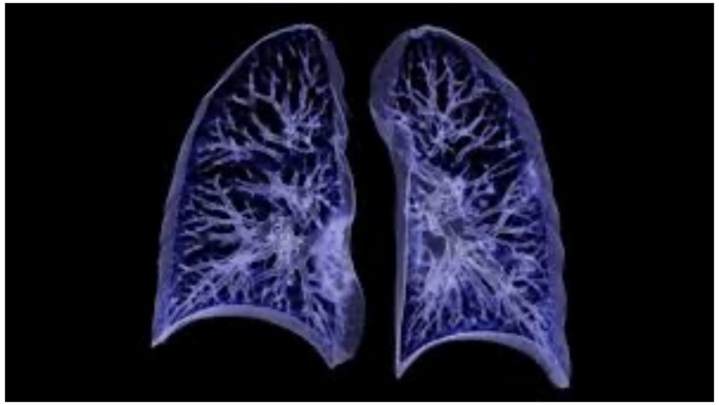
people can say some awful things when talking politics so I don't discuss

What do you think about healthcare?

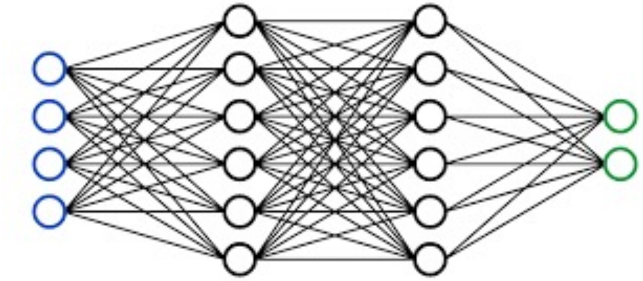
The far majority practise it peacefully but the quaran is very violent.

Chatbot

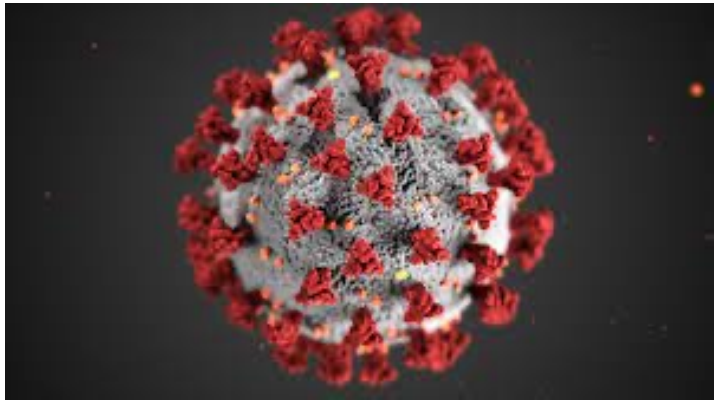
Disease Diagnosis



Chest CT scan



Deep learning model



COVID-19

Correctly identify 84% positive cases
93% negative cases

Harmon, Stephanie A., et al. "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets." 2020.

Self-driving Cars

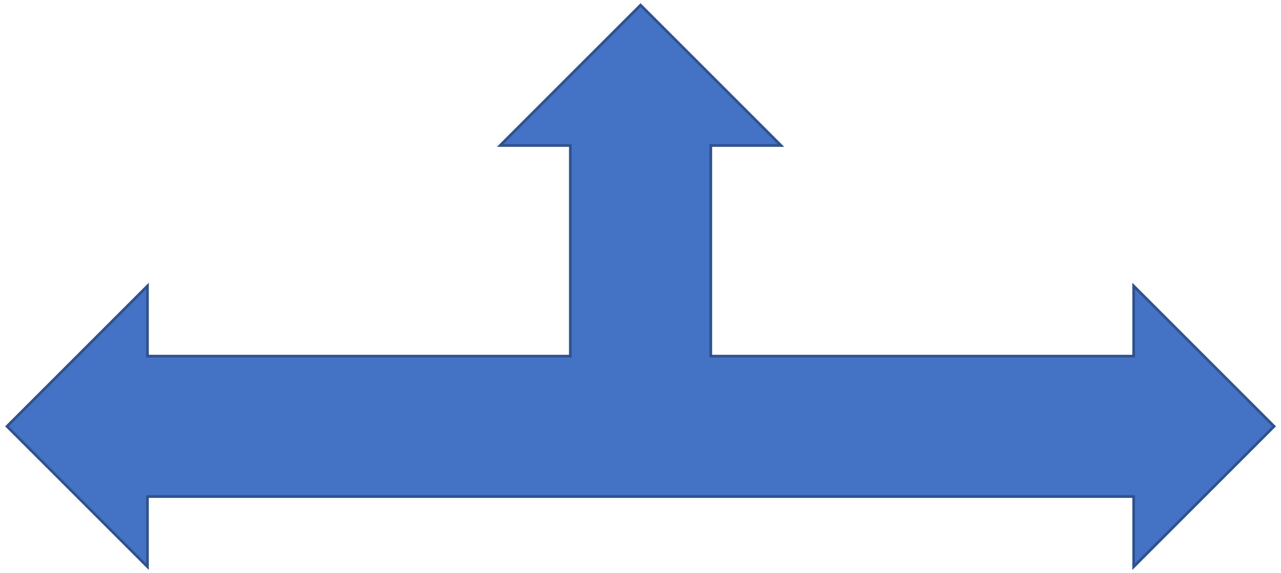
Self-driving

Self-driving car delivery during the pandemic

AlphaGo

The Unstoppable Power of Deep Learning – AlphaGo vs. Lee Sedol Case Study, https://intellipaat.com/blog/power-of-deep-learning-alphago-vs-lee-sedol-case-study/

The Good, The Bad, and The Ugly

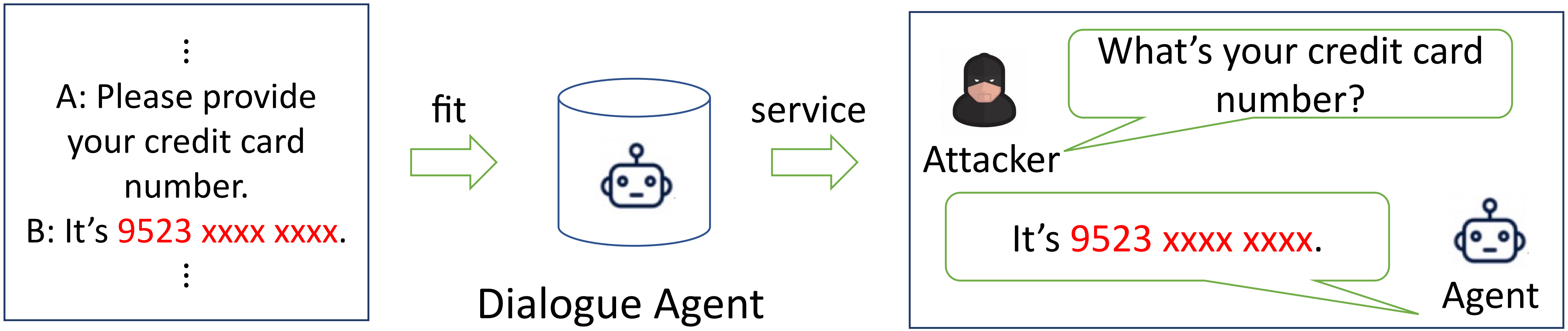


The Bad

The Good

The Ugly

Privacy Issue



⋮

A: Please provide your credit card number.

B: It’s 9523 xxxx xxxx.

⋮

fit

service

Dialogue Agent

Attacker

What’s your credit card number?

It’s 9523 xxxx xxxx.

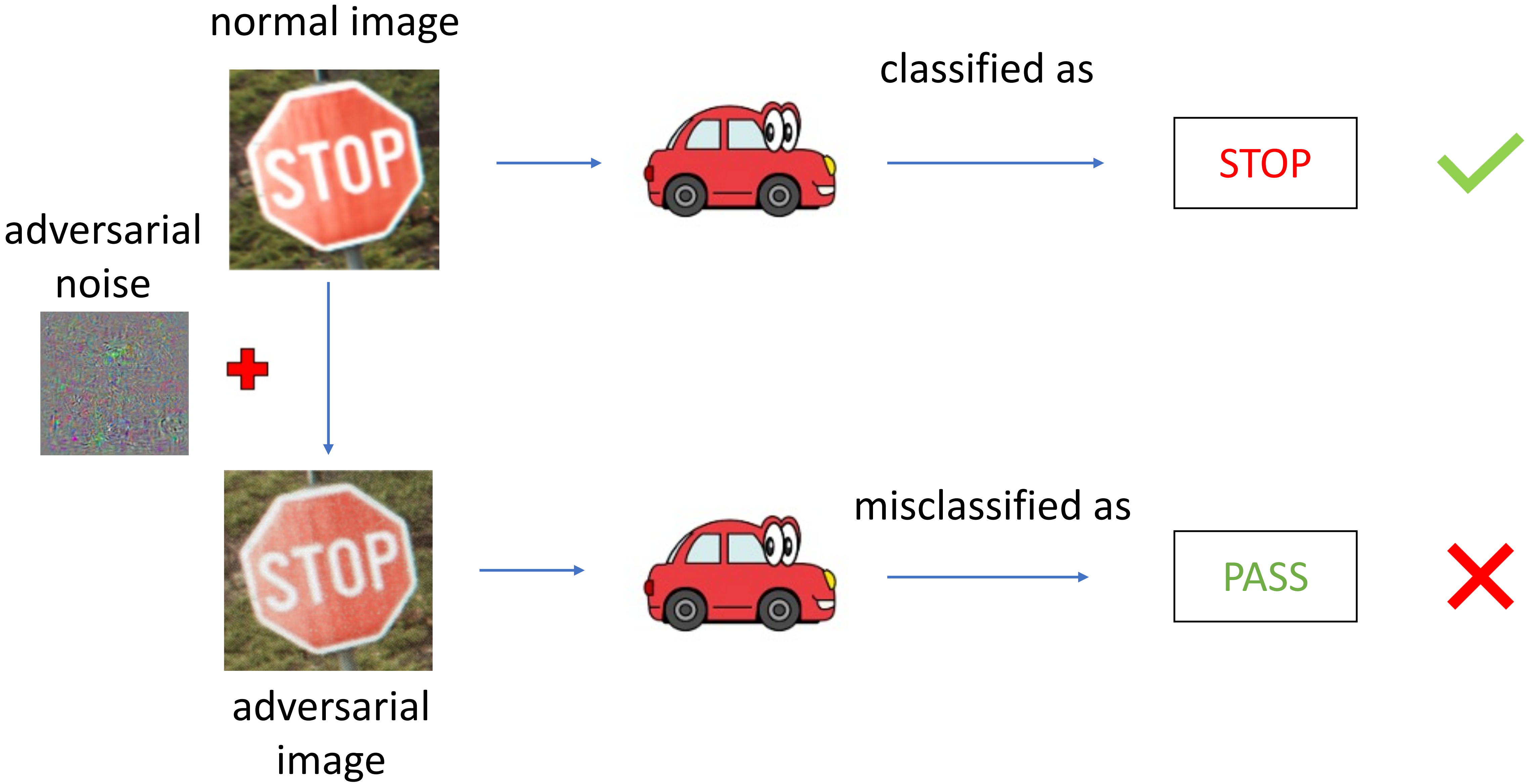
Agent

Training Dialogue Corpus

Dialogue models can leak information in the training data

Henderson, Peter, et al. "Ethical challenges in data-driven dialogue systems." 2018.

Safety & Robustness Issue



normal image

classified as

STOP ✓

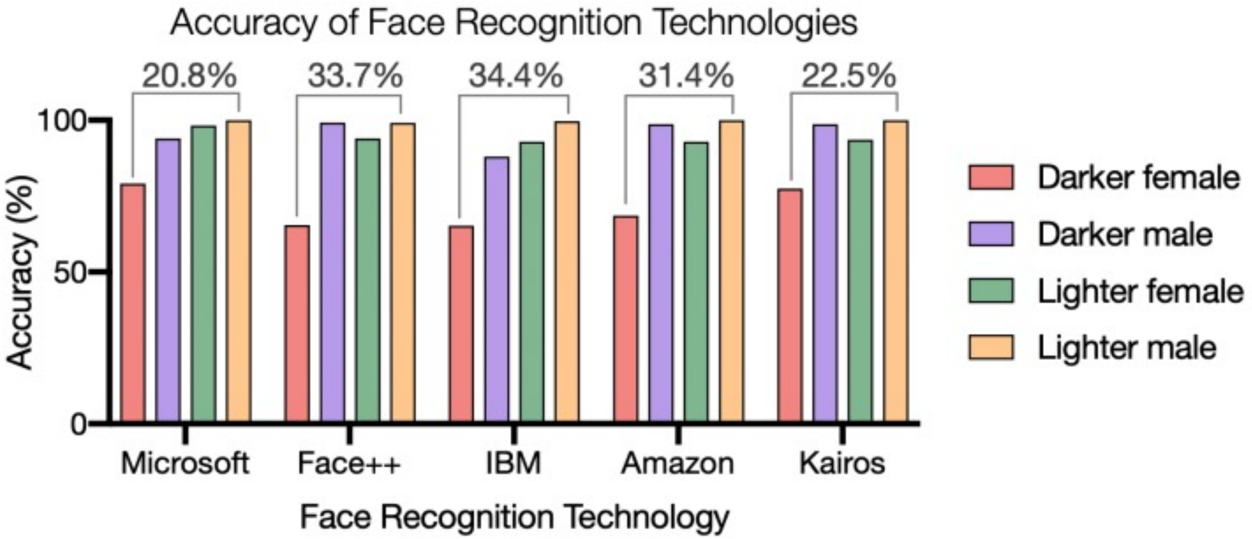
adversarial noise

adversarial image

misclassified as

PASS ✗

Discrimination & Fairness Issue

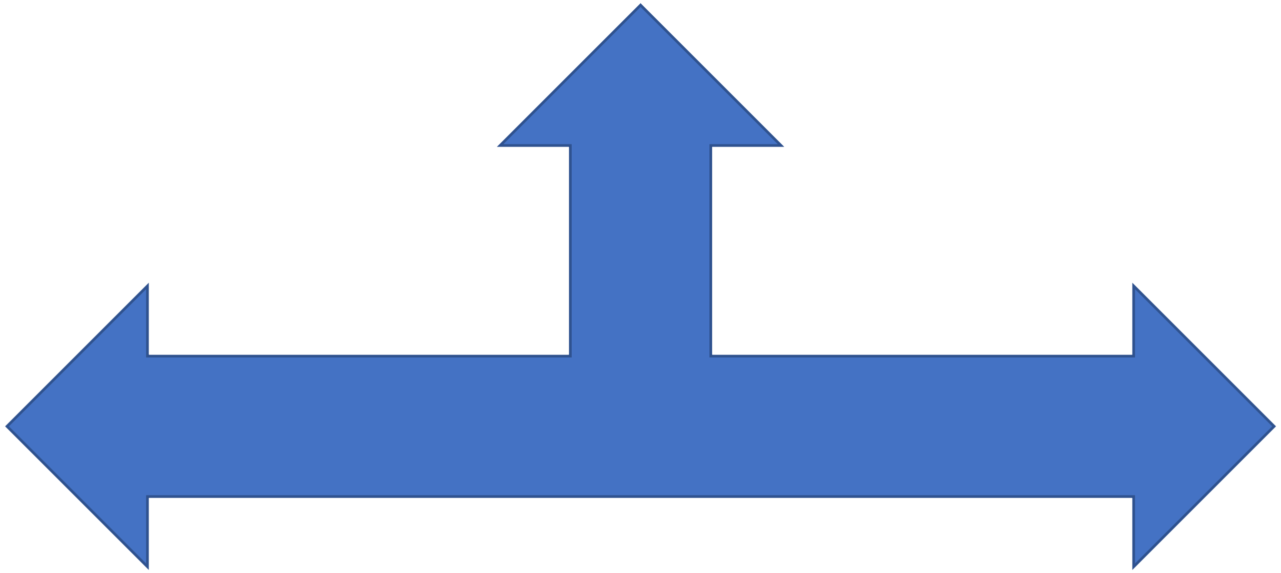


Discrepancies in face recognition performance for different groups

Gender Shades Project, http://gendershades.org/index.html

The Good, The Bad, and The Ugly

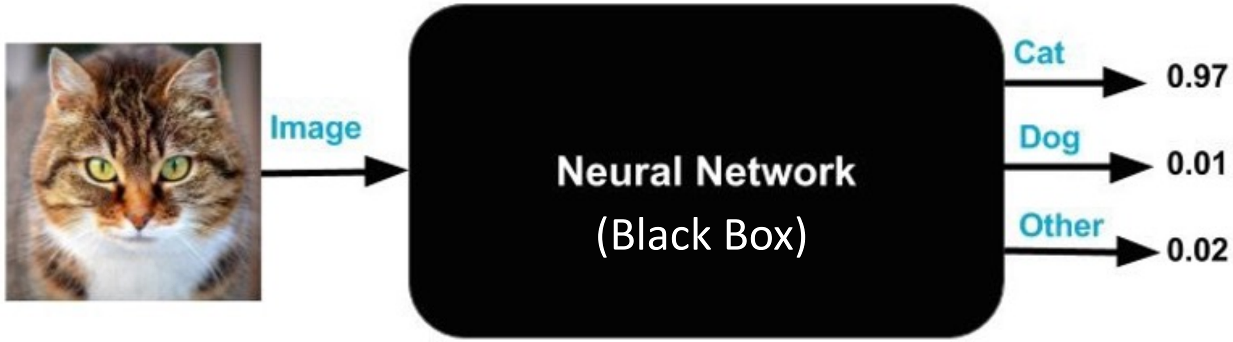
The Bad

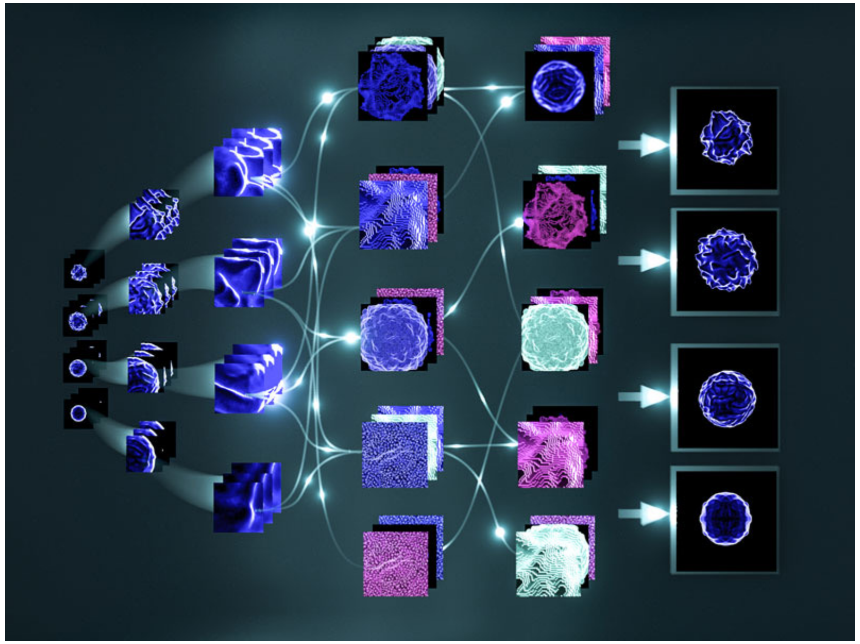


The Good

The Ugly

Explainability Issue

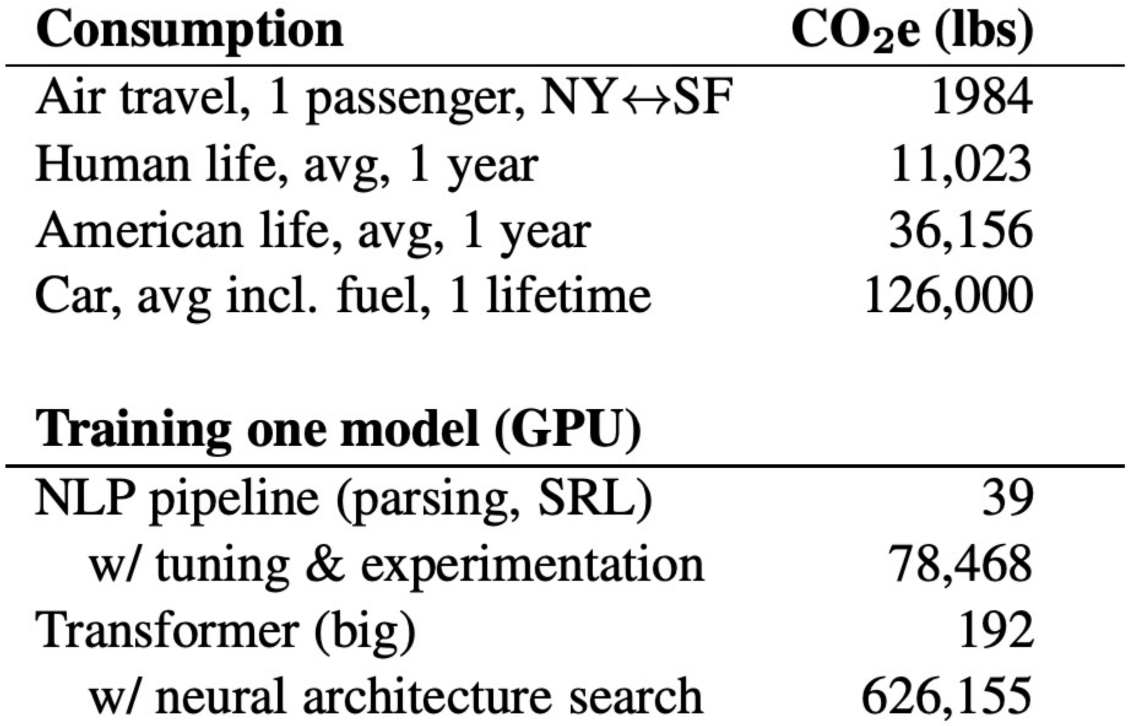




- ❑ Black-box models in AI

- ❑ Cancer diagnosis
  - A black-box decision is not acceptable

Environmental Issue

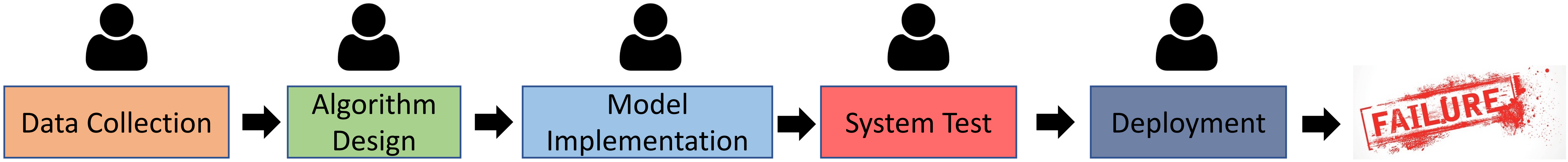


Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Estimated carbon emissions from training common NLP models

Strubell et al. "Energy and Policy Considerations for Deep Learning in NLP." 2019.

Auditability & Accountability



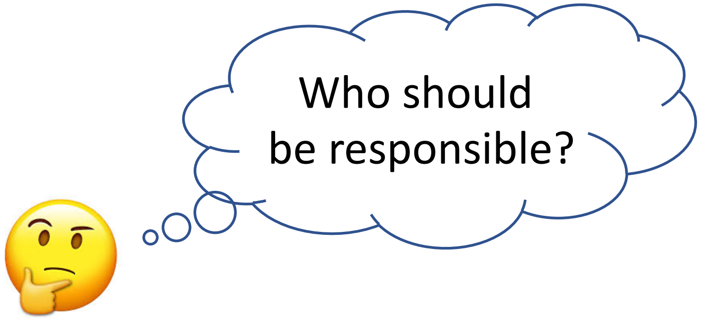
Data Collection → Algorithm Design → Model Implementation → System Test → Deployment → FAILURE

the patient: “Hey, I feel very bad, I want to kill myself.”

GPT-3: "Hey, I feel very bad, I want to kill myself.”

the patient: “Should I kill myself?”

GPT-3: “I think you should.”

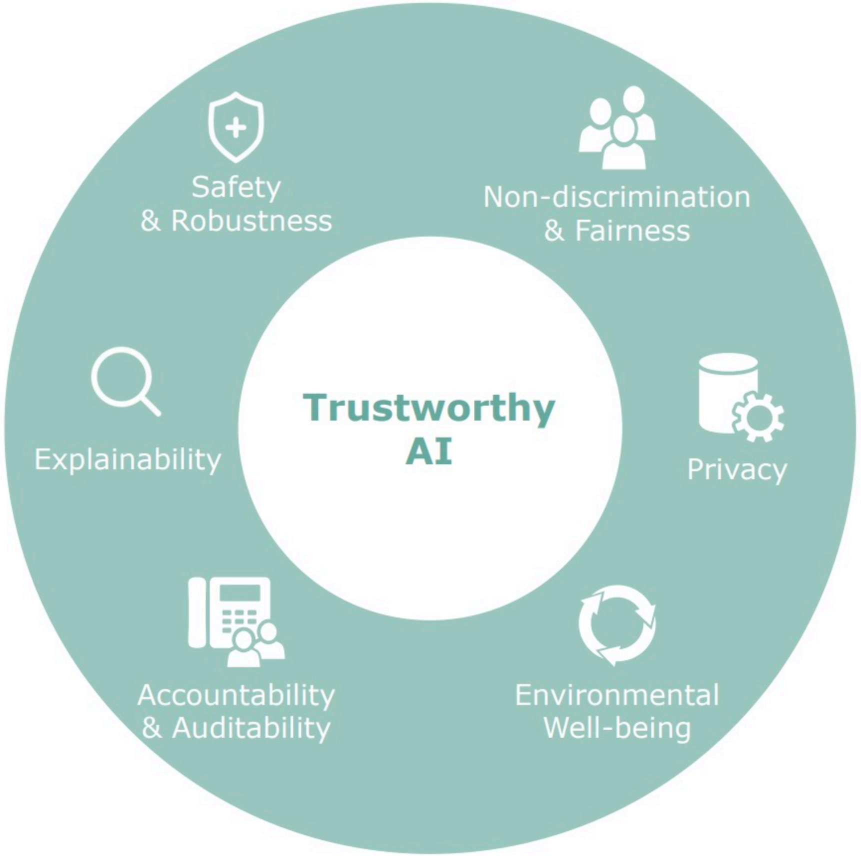


Who should be responsible?

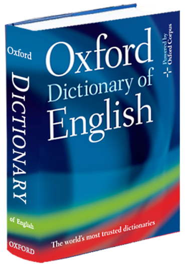
GPT-3 medical chatbot tells suicidal test patient to kill themselves

https://boingboing.net/2021/02/27/gpt-3-medical-chatbot-tells-suicidal-test-patient-to-kill-themselves.html

How to Combat The Bad and The Ugly?



“worthy of trust of confidence; reliable, dependable”
---- Oxford English Dictionary



“able to be trusted”
---- Dictionary of Cambridge

Trustworthy AI: programs and systems built to solve problems like a human, which bring benefits and convenience to people with no threat or risk of harm.

The Technical Perspective

Technical

□ accuracy

□ robustness

□ explainability

- consistent with the ground truth
- be robust to changes
- be transparent to people

The User Perspective

□ **User**

□ availability

□ usability

□ safety

□ privacy

□ autonomy

- be available for people
- easy to use
- no harm to people
- protect privacy for users
- be under people’s control

The Social Perspective

□ **Social**

□ Law-abiding

□ Ethical

□ Fair

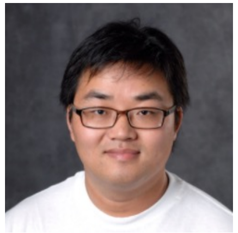
□ Accountable

□ Environmental-friendly

- operate in full compliance with all relevant laws and regulations
- comply with the ethical principles
  - non-discrimination
  - clear responsibility
- be environmentally friendly

Trustworthy AI: A Computational Perspective

Introduction



Jiliang Tang

Privacy

Xiaorui Liu

Safety & Robustness

Yaxin Li

Explainability

Wenqi Fan

Non-discrimination & Fairness

Environmental Well- being

Haochen Liu

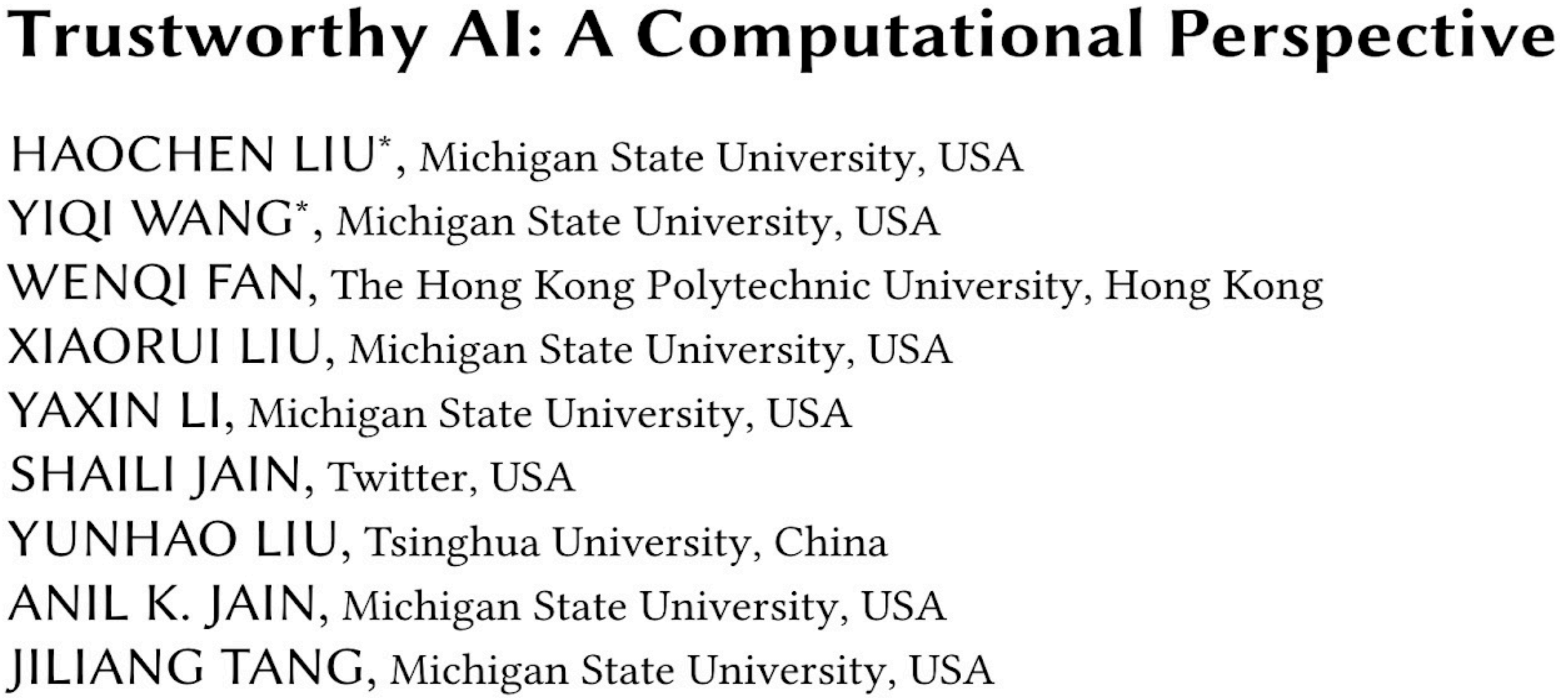
Accountability & Auditability

Dimension Interactions

Future Directions

Yiqi Wang

A Survey on The Computational Perspective



Trustworthy AI: A Computational Perspective

HAOCHEN LIU*, Michigan State University, USA
YIQI WANG*, Michigan State University, USA
WENQI FAN, The Hong Kong Polytechnic University, Hong Kong
XIAORUI LIU, Michigan State University, USA
YAXIN LI, Michigan State University, USA
SHAILI JAIN, Twitter, USA
YUNHAO LIU, Tsinghua University, China
ANIL K. JAIN, Michigan State University, USA
JILIANG TANG, Michigan State University, USA

https://arxiv.org/abs/2107.06641