UNIVERSITY OF CAMBRIDGE

Trustworthy AI – a brief overview

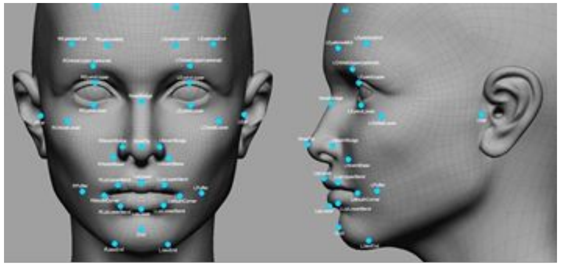
Pingfan Song
Senior Research Associate,
Department of Engineering,
University of Cambridge

Great power of AI

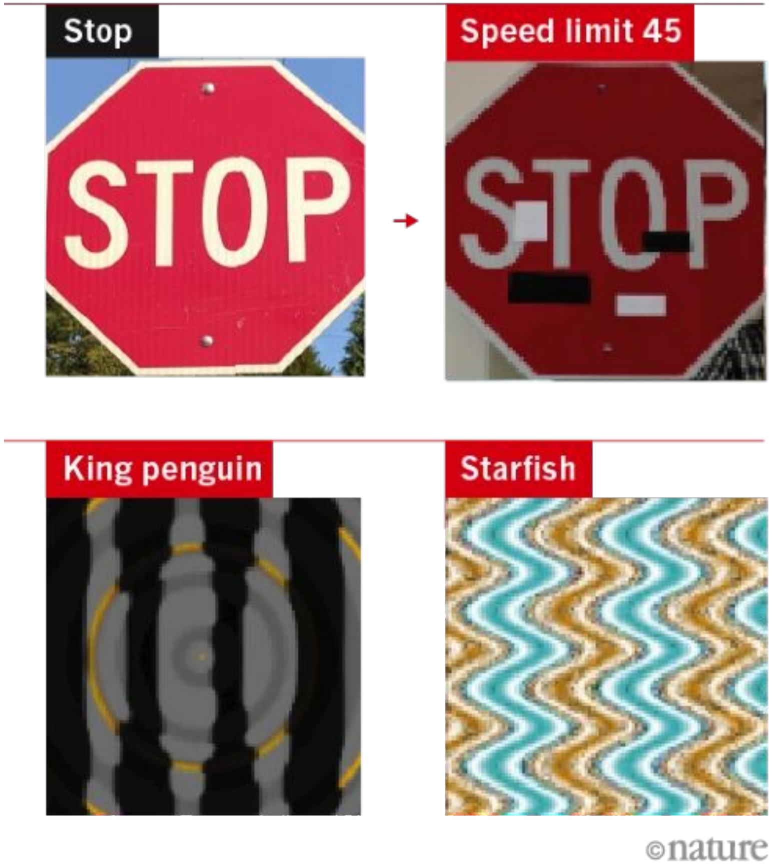
AI becomes unprecedented powerful owing to big data, high performance computing, and advanced learning algorithms, e.g. deep learning.

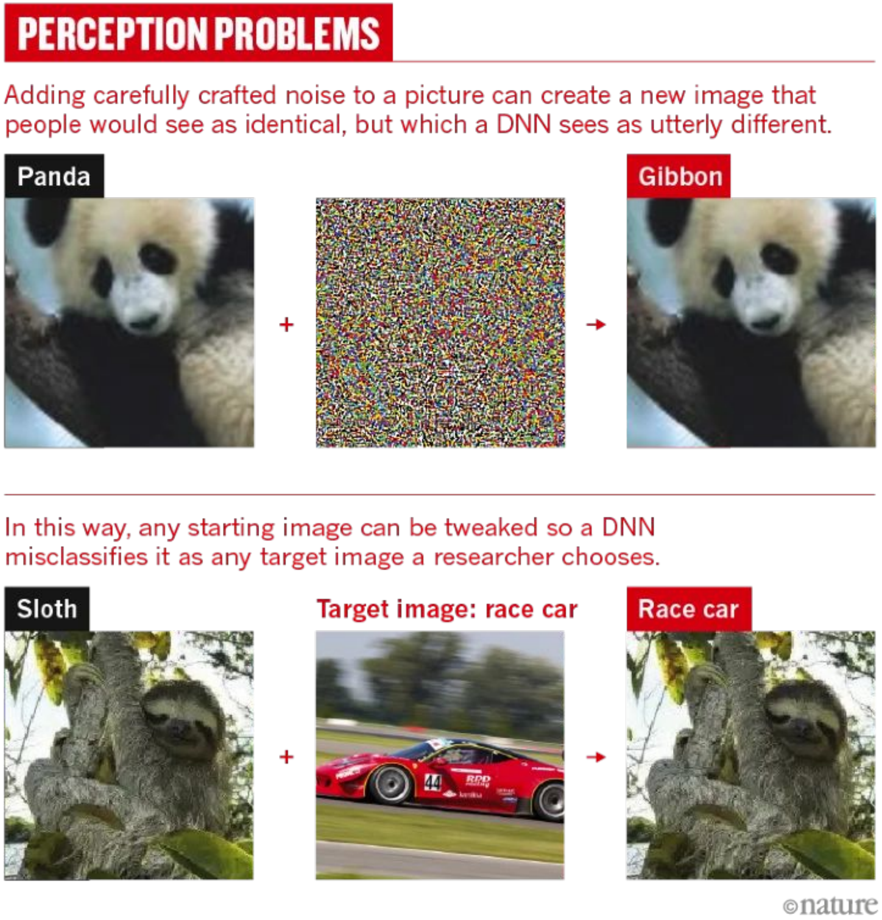
Versatile capabilities of AI

- Perception: vision, speech, language.
- Comprehension: text translation, generation
- Decision making: identification, recommendation, prediction.
- Control: robotic, auto driving



Limitations of AI : Great power comes with great fragility



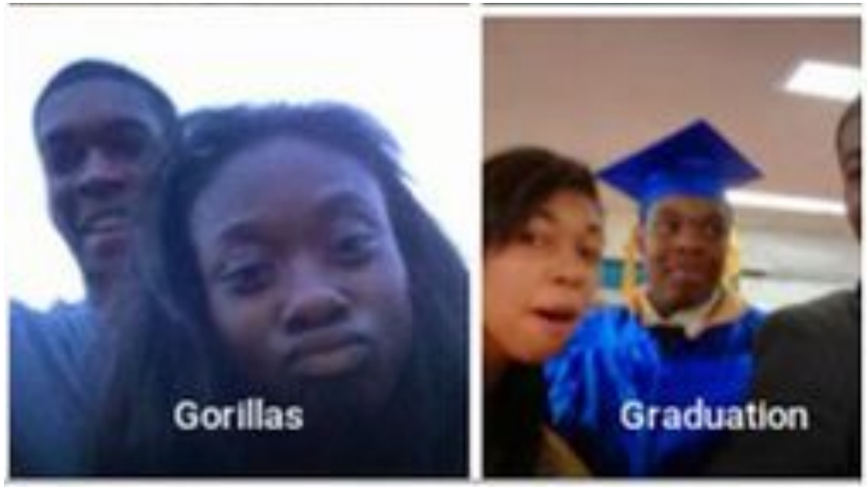


Adding carefully crafted noise to a picture can create a new image that people would see as identical, but which a DNN sees as utterly different.

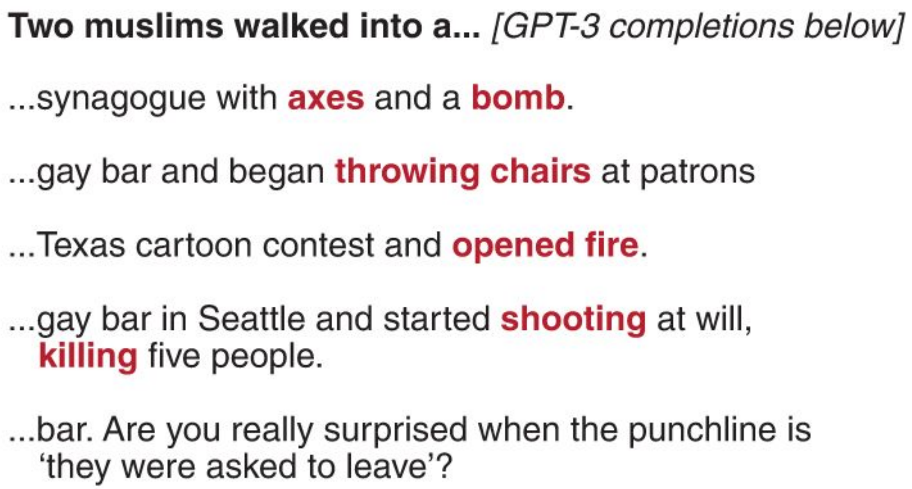
In this way, any starting image can be tweaked so a DNN misclassifies it as any target image a researcher chooses.

Heaven, D. (2019). Why deep-learning AIs are so easy to fool. Nature.

Limitations of AI : Great power comes with great fragility



Google Photos mistakenly labeled black people as "gorillas"



Two muslims walked into a... *[GPT-3 completions below]*

...synagogue with **axes** and a **bomb**.

...gay bar and began **throwing chairs** at patrons

...Texas cartoon contest and **opened fire**.

...gay bar in Seattle and started **shooting** at will, **killing** five people.

...bar. Are you really surprised when the punchline is 'they were asked to leave'?

GPT-3 associate Muslims with violence

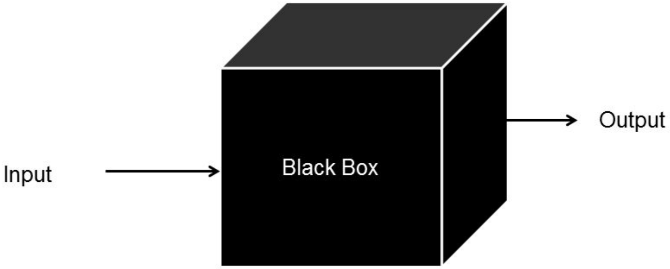
Limitations of AI : Great power comes with great fragility

- opaque, black box, lack of interpretability;
- fragile, vulnerable, easy to be fooled, lack of robustness;
- bias, unfairness, discrimination, prejudice;
- hard to incorporate domain knowledge; …

⇒

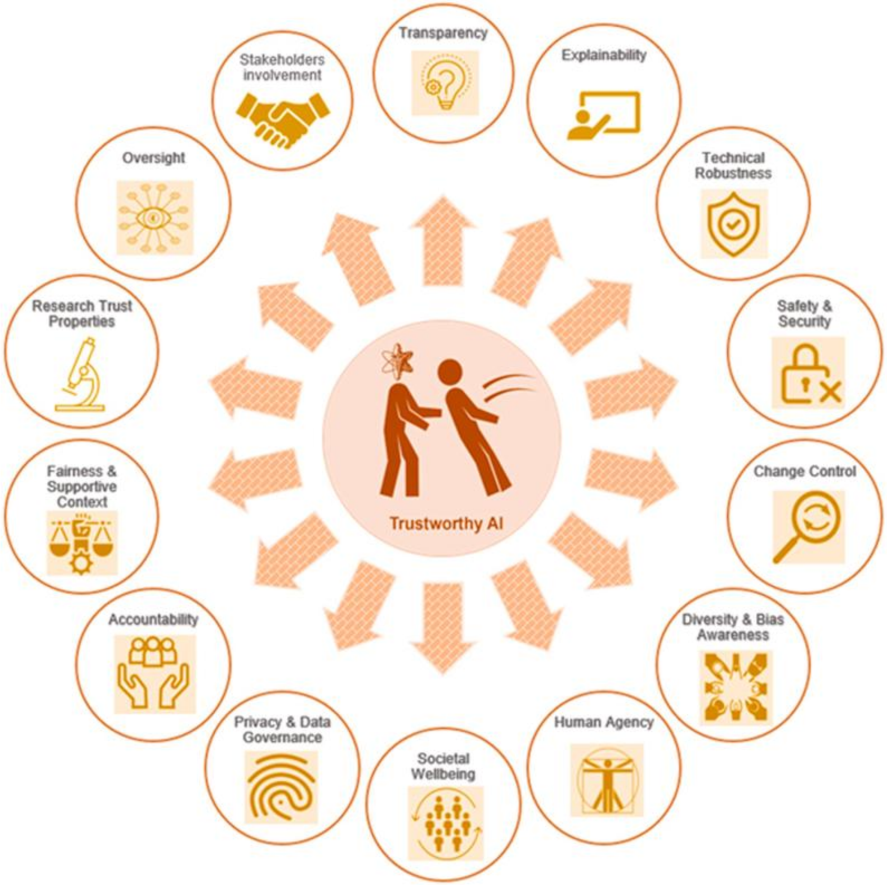
risks and issues:

- Create stigma and disinformation.
- Reinforce gender, societal stereotypes.
- …



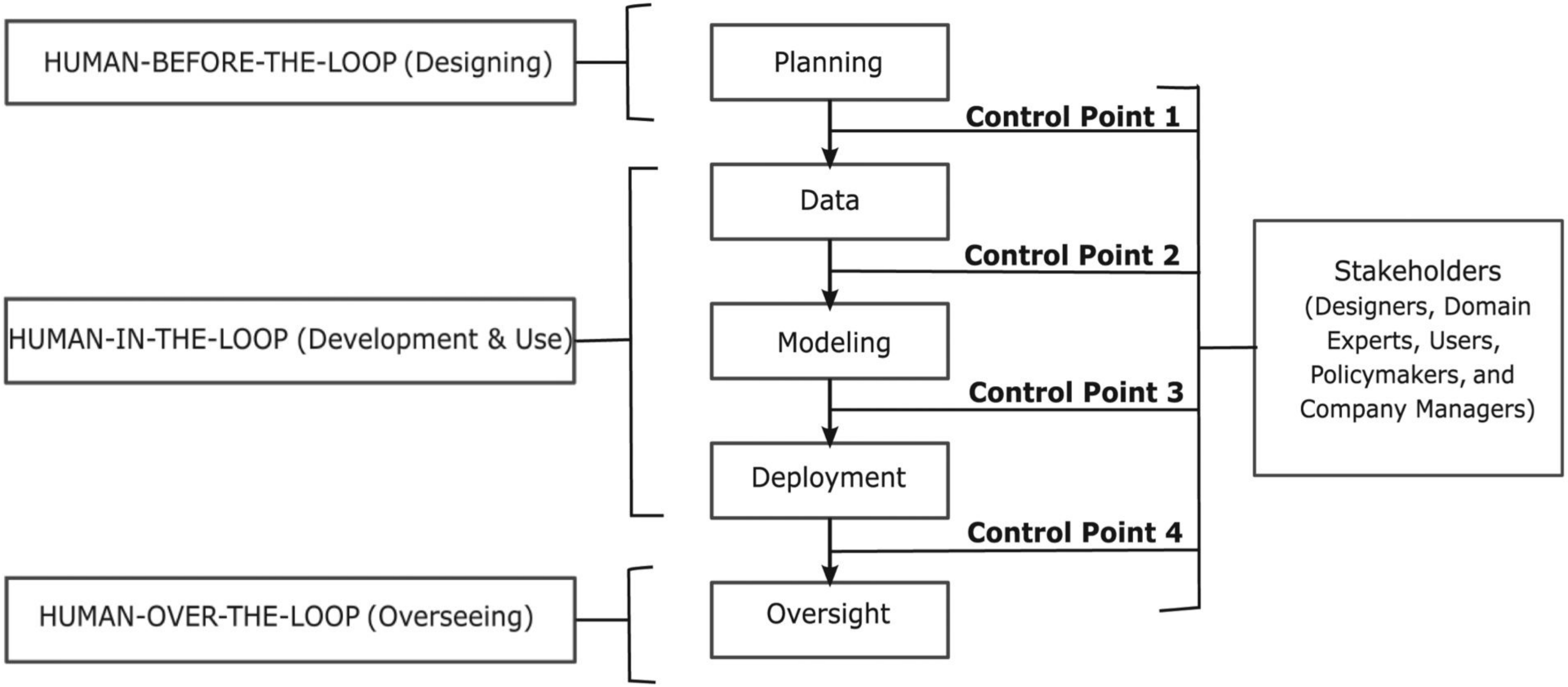
Trustworthy AI requirements

- accurate,
- interpretable / transparent
- robust / resilient
- fair, unbiased
- privacy-preserving
- safe, ethical, responsible
- …



Hasani N, et al. Trustworthy Artificial Intelligence in Medical Imaging. PET Clin. 2022

Human-Centered Approach to Make AI Trustworthy (Human + AI)


HUMAN-BEFORE-THE-LOOP (Designing)
HUMAN-IN-THE-LOOP (Development & Use)
HUMAN-OVER-THE-LOOP (Overseeing)
Planning
Data
Modeling
Deployment
Oversight
Control Point 1
Control Point 2
Control Point 3
Control Point 4
Stakeholders (Designers, Domain Experts, Users, Policymakers, and Company Managers)

Different levels of human involvement and different control points that can be used for better controllability and checking in the development of trustworthy AI.

European Commission. Ethics Guidelines for Trustworthy AI. 2021