ERCIM Forum on Digital Ethics in Research
Paris,

October 17/18, 2022

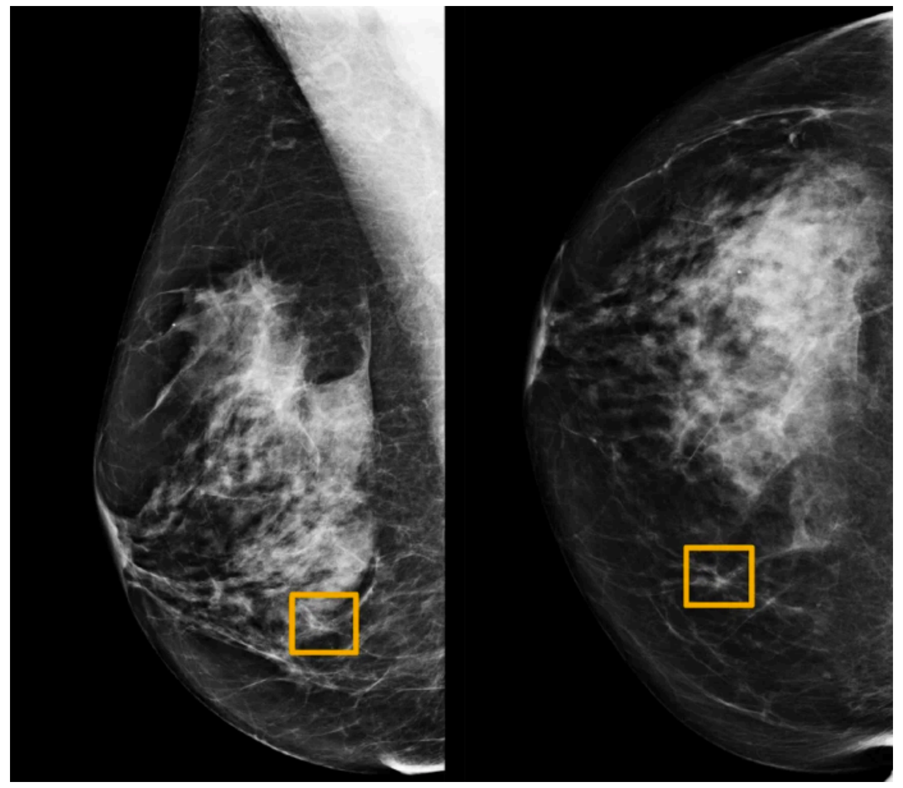
# Responsible Development, Use and Governance of Artificial Intelligence

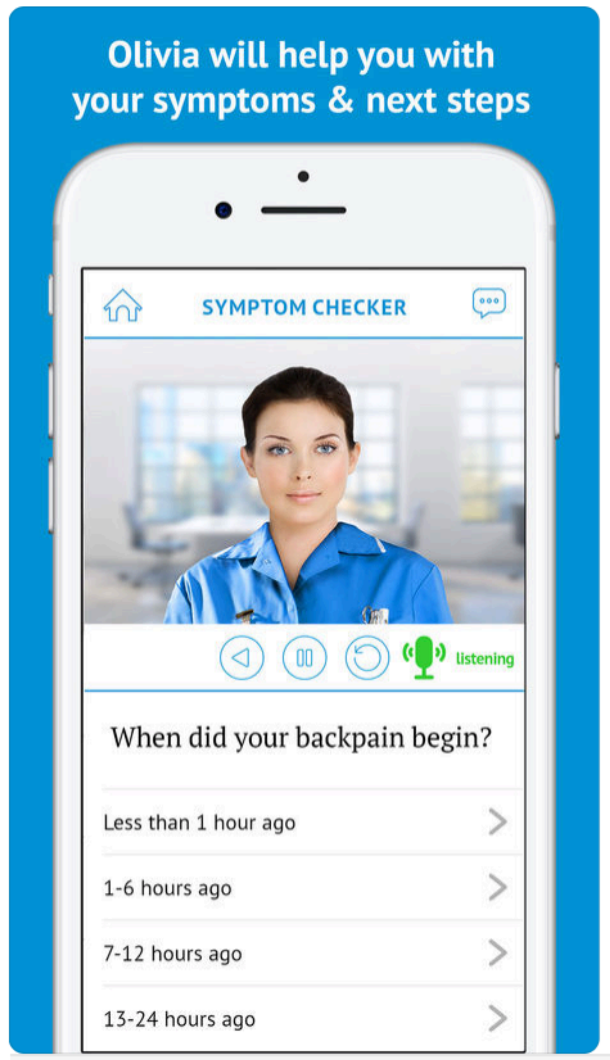
Raja Chatila
Institute of Intelligent Systems and Robotics (ISIR)
Faculty of Sciences and Engineering, Pierre and Marie Curie Campus
Sorbonne University, Paris, France

Raja.Chatila@sorbonne-universite.fr

# Multiple Applications of AI And Robotics

- Transportation, logistics, delivery
- Healthcare
- Manufacturing
- Agriculture
- Personal services & assistance
- Security
- Recommender systems, advertisement
- Recruitment & management
- Insurance & finance
- Justice
- Warfare
- …



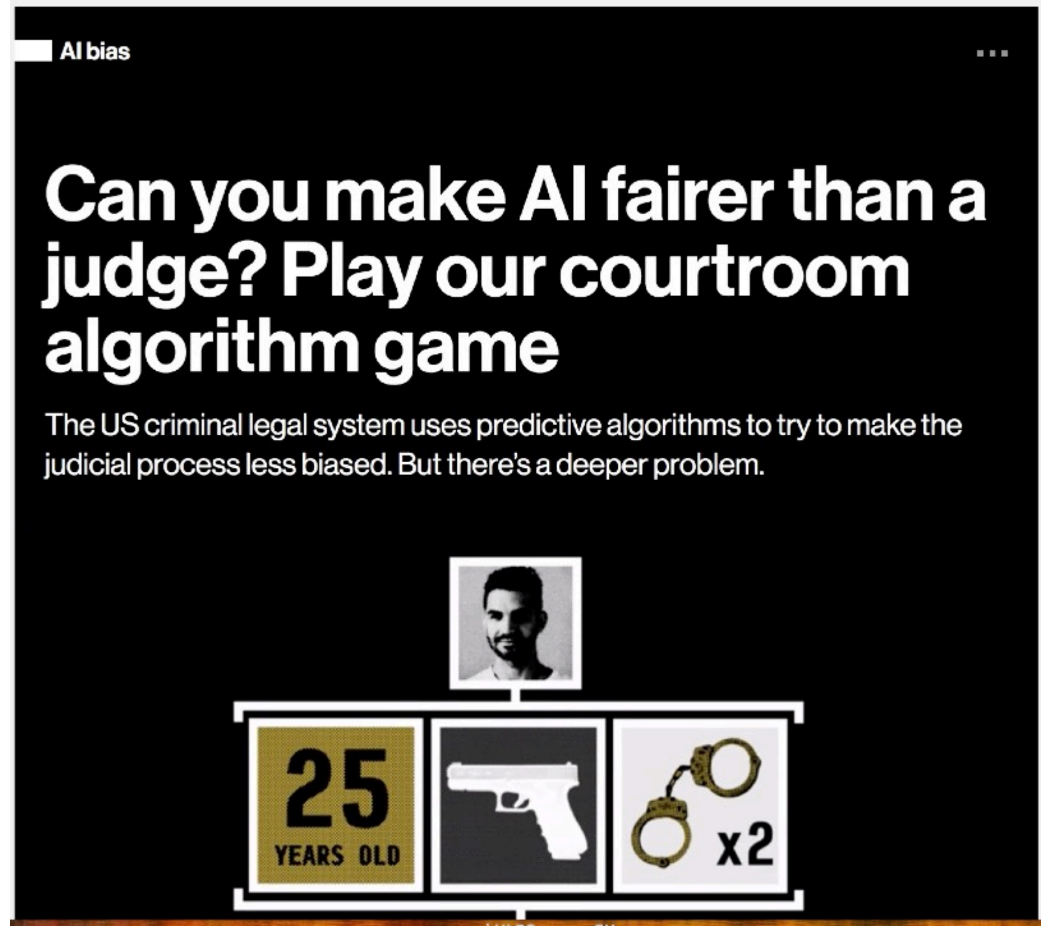


Olivia will help you with your symptoms & next steps

SYMPTOM CHECKER

When did your backpain begin?

- Less than 1 hour ago
- 1-6 hours ago
- 7-12 hours ago
- 13-24 hours ago



AI bias

Can you make AI fairer than a judge? Play our courtroom algorithm game

The US criminal legal system uses predictive algorithms to try to make the judicial process less biased. But there's a deeper problem.

**A face-scanning algorithm increasingly decides whether you deserve the job**

HireVue claims it uses **artificial intelligence** to decide who's best for a job. Outside experts call it 'profoundly disturbing.'

# What is an Computational "Intelligent" System?

- A computational intelligent system is a set of **algorithms designed by humans**, using data (big/small/sensed) to solve [more or less] complex problems in [more or less] complex situations.

- The system might include deductive inference, as well as machine learning processes, *i.e.*, the capability of improving its performance based on data classification to build **statistical models** from data (*e.g.*, deep learning), or on evaluating previous decisions (*e.g.*, reinforcement learning).

- Such systems could be regarded as "autonomous" in a **given domain** and for **specific tasks**, as long as they are capable of accomplishing these tasks despite environment variations within this domain.

- Difference between automated and autonomous systems is related to **complexity** of task and domain, and **importance** of variations

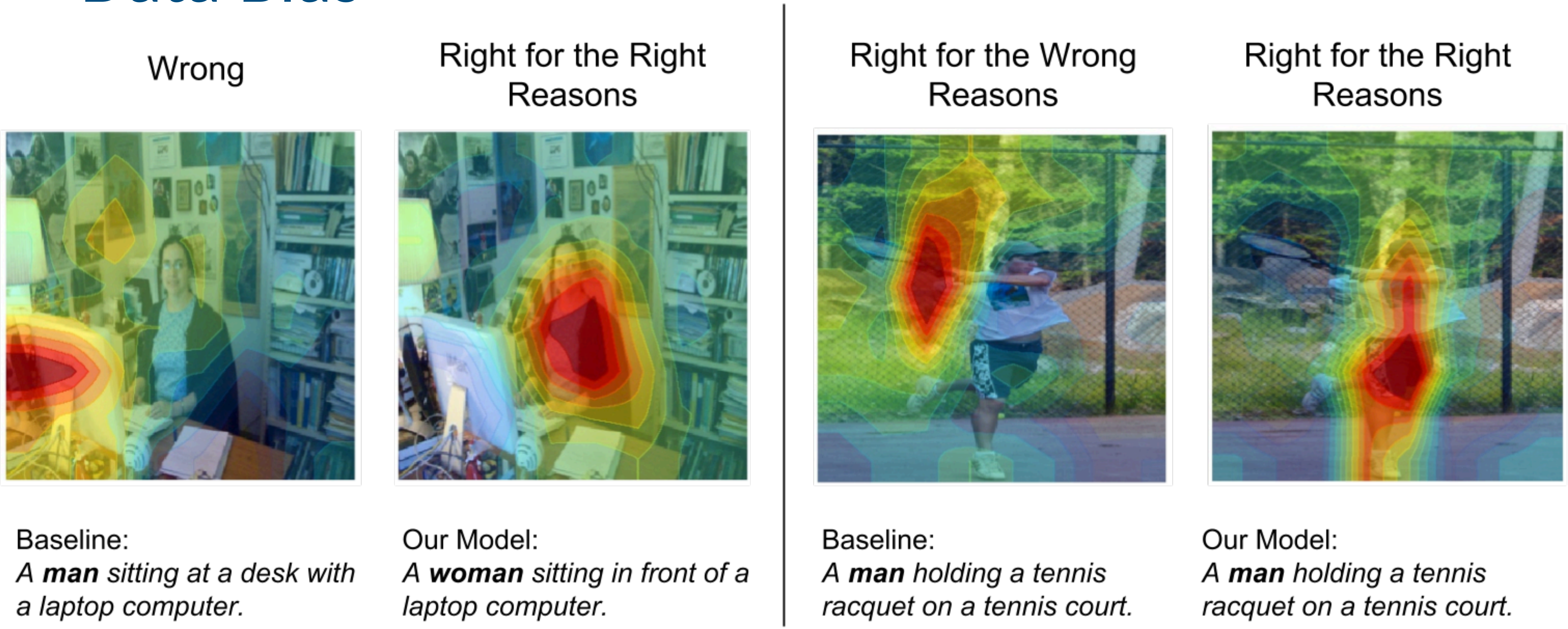
# Machine Learning

Statistical data processing and classification

- Use of probability distributions, correlations, …
- Use of artificial neural nets as classifiers
- Optimization algorithms

- Supervised learning: correct answer provided by a truth model.
- Unsupervised learning: search for regularities in the data
- Reinforcement Learning: select the most promising action based on rewards

# Machine Learning Limitations
# Data Bias



Wrong

Right for the Right Reasons

Right for the Wrong Reasons

Right for the Right Reasons

Baseline:
*A **man** sitting at a desk with a laptop computer.*

Our Model:
*A **woman** sitting in front of a laptop computer.*

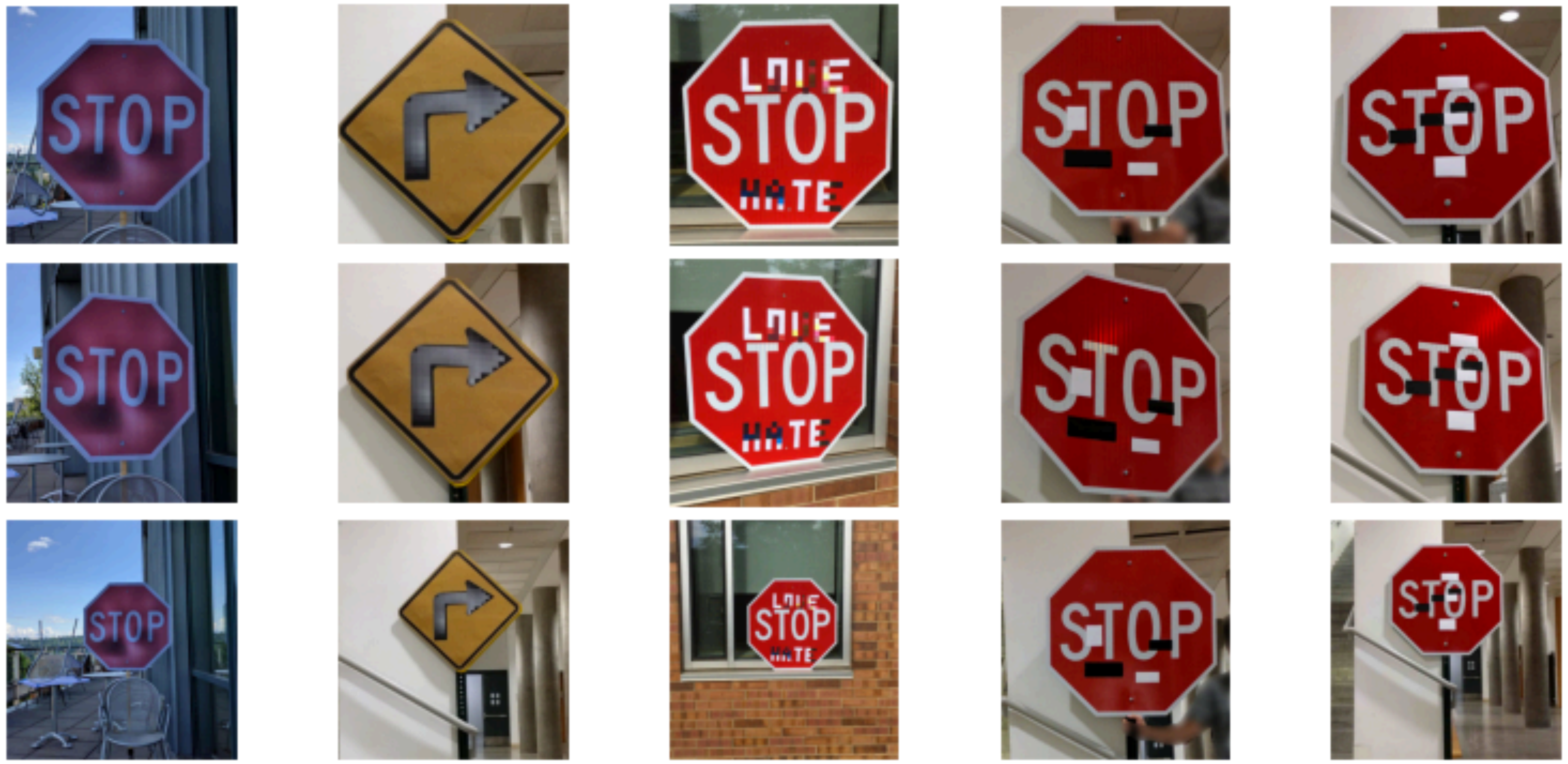
Baseline:
*A **man** holding a tennis racquet on a tennis court.*

Our Model:
*A **man** holding a tennis racquet on a tennis court.*

*Women also Snowboard: Overcoming Bias in Captioning Models.*

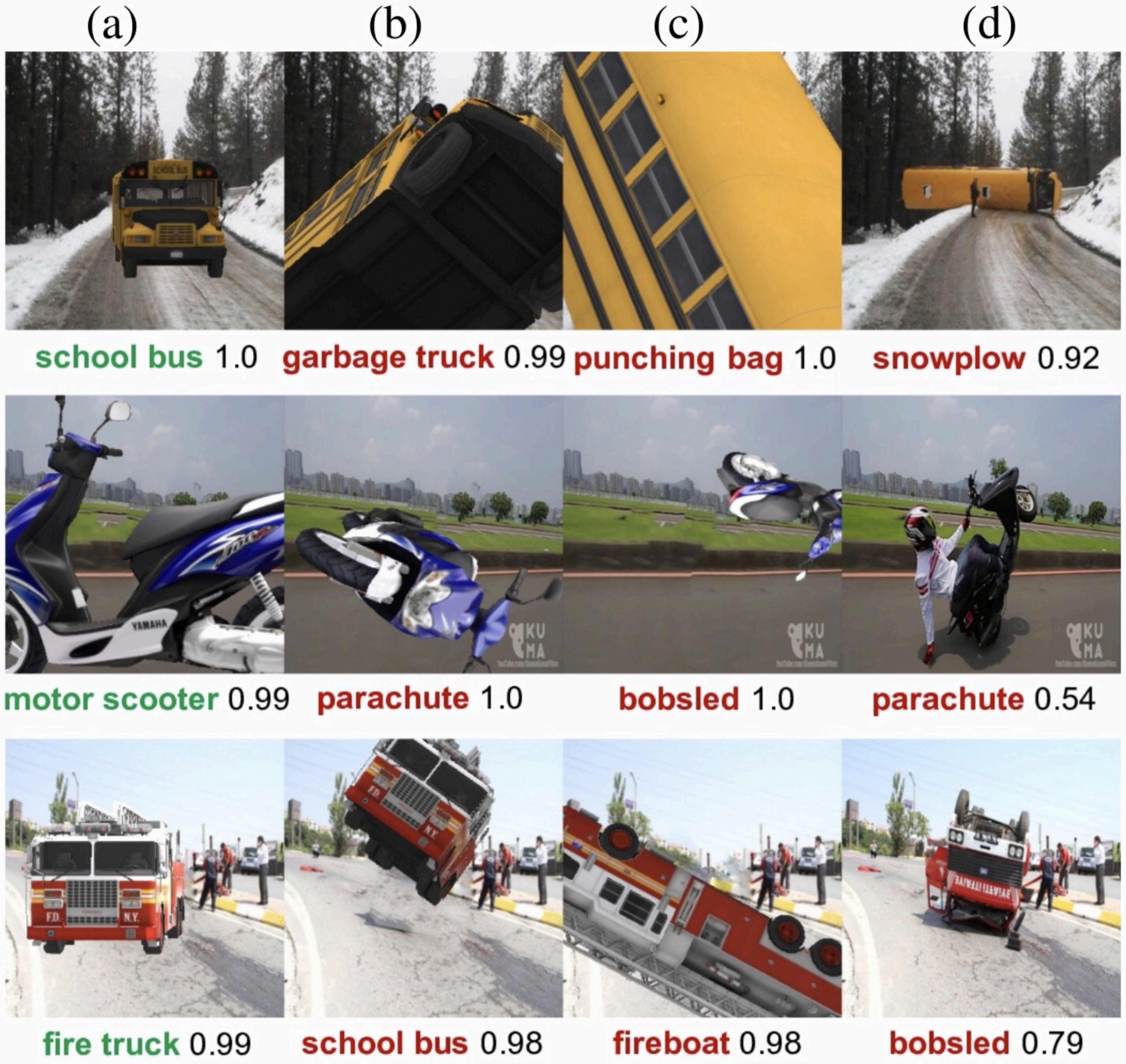
Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, Anna Rohrbach. ECCV 2018

# Deep Learning Limitations Robustness



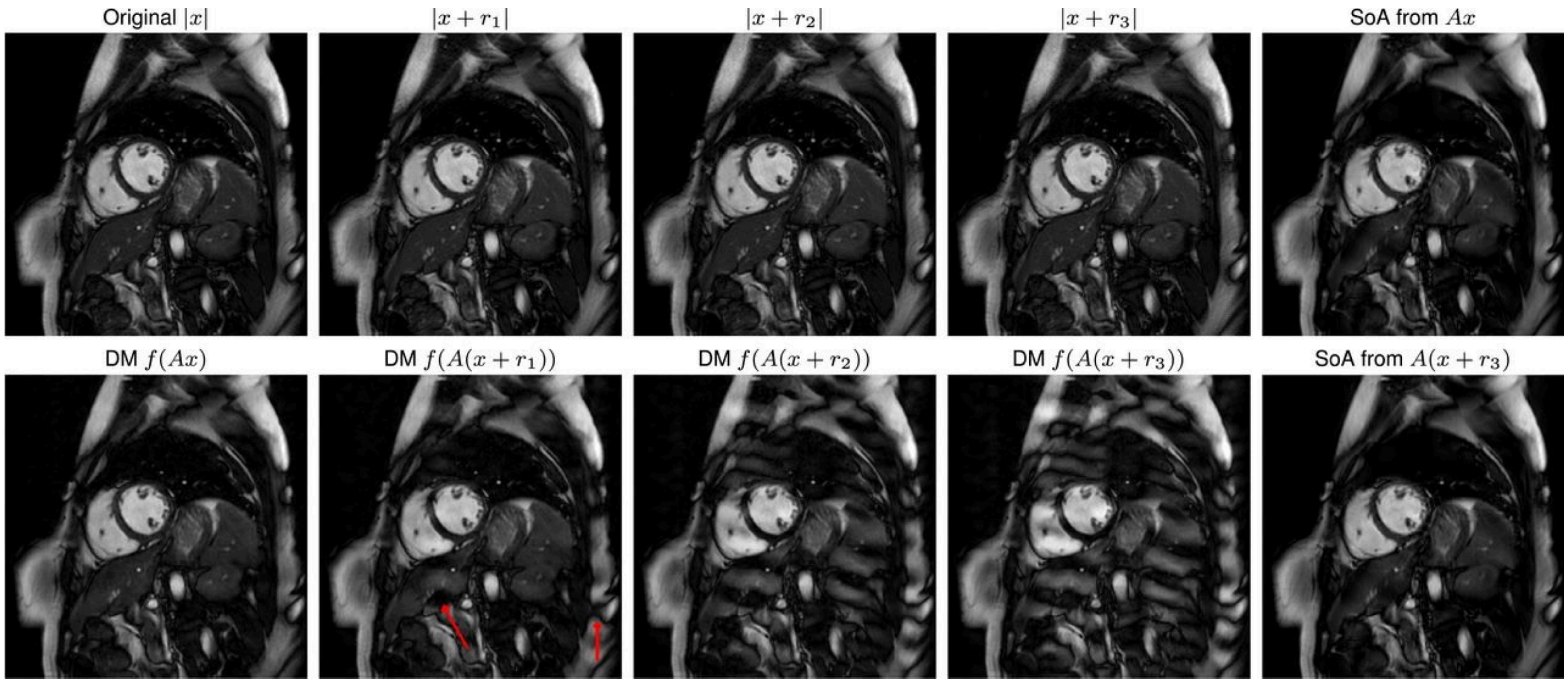
Targeted physical perturbation experiment
The misclassification target was Speed Limit 45.

Robust Physical-World Attacks on Deep Learning Models K. Eykholt et al. CVPR 2018.



Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. Michael A. Alcorn et al., CVPR 2019

# Deep Learning Limitations
## Instabilities



Perturbations $r_j$ (created to simulate worst-case effect) with $|r_1|<|r_2|<|r_3|$ are added to the image x.

(*Top*) Images 1 to 4 are original image x and perturbations $x+r_j$.

(Bottom) Images 1 to 4 are reconstructions from $A(x+r_j)$ using the Deep MRI (DM) network f, where A is a subsampled Fourier transform (33% subsampling);

(*Top* and *Bottom*) Image 5 is a reconstruction from Ax and $A(x+r_3)$ using an SoA method; Note how the artifacts (red arrows) are hard to dismiss as nonphysical.

**On instabilities of deep learning in image reconstruction and the potential costs of AI**

Vegard Antun et al. PNAS 2020;117:48:30088-30095

# Machine Learning capabilities
# Generative Adversarial Networks (GAN)

- Two networks, on producing “fake” data, the other trying to classify real from fake.
- The two networks simultaneously learn to improve.
- GANs can learn and imitate any data distribution.

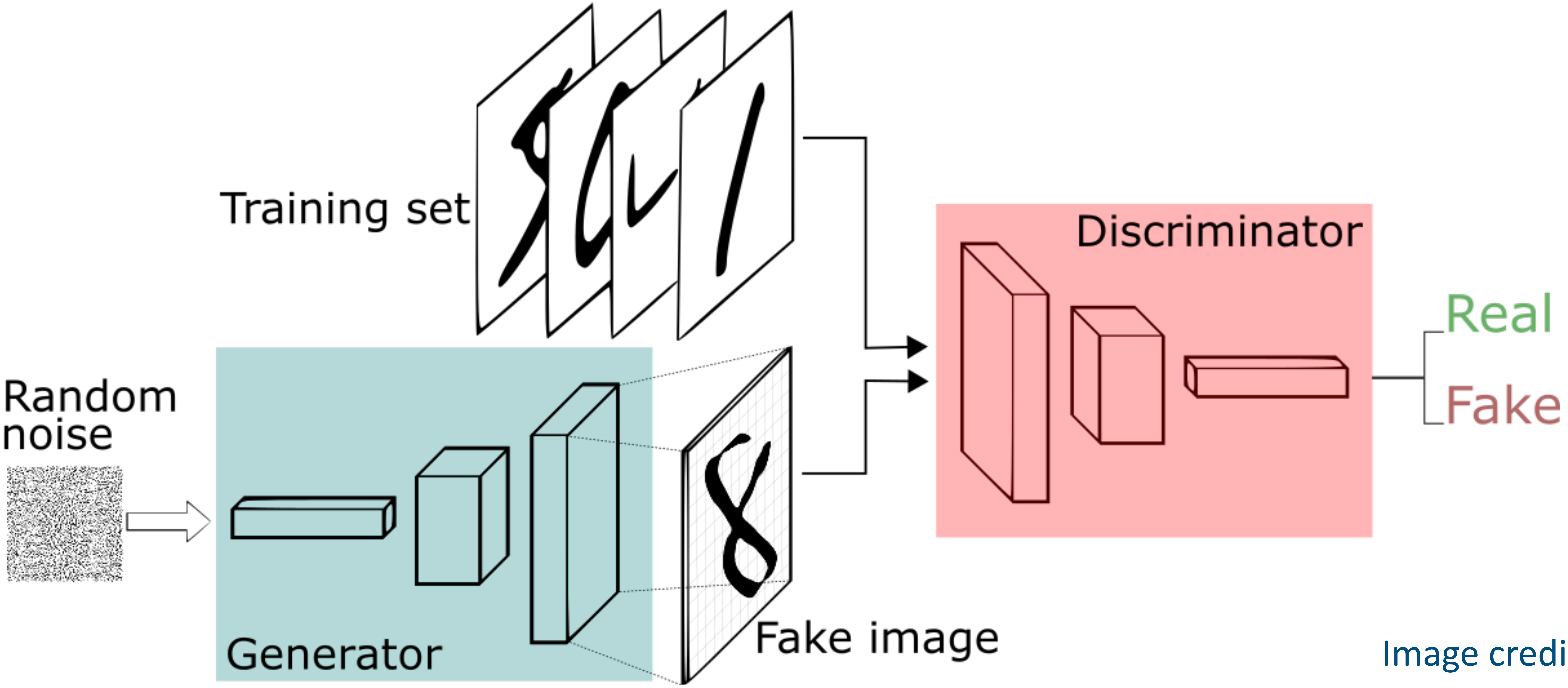


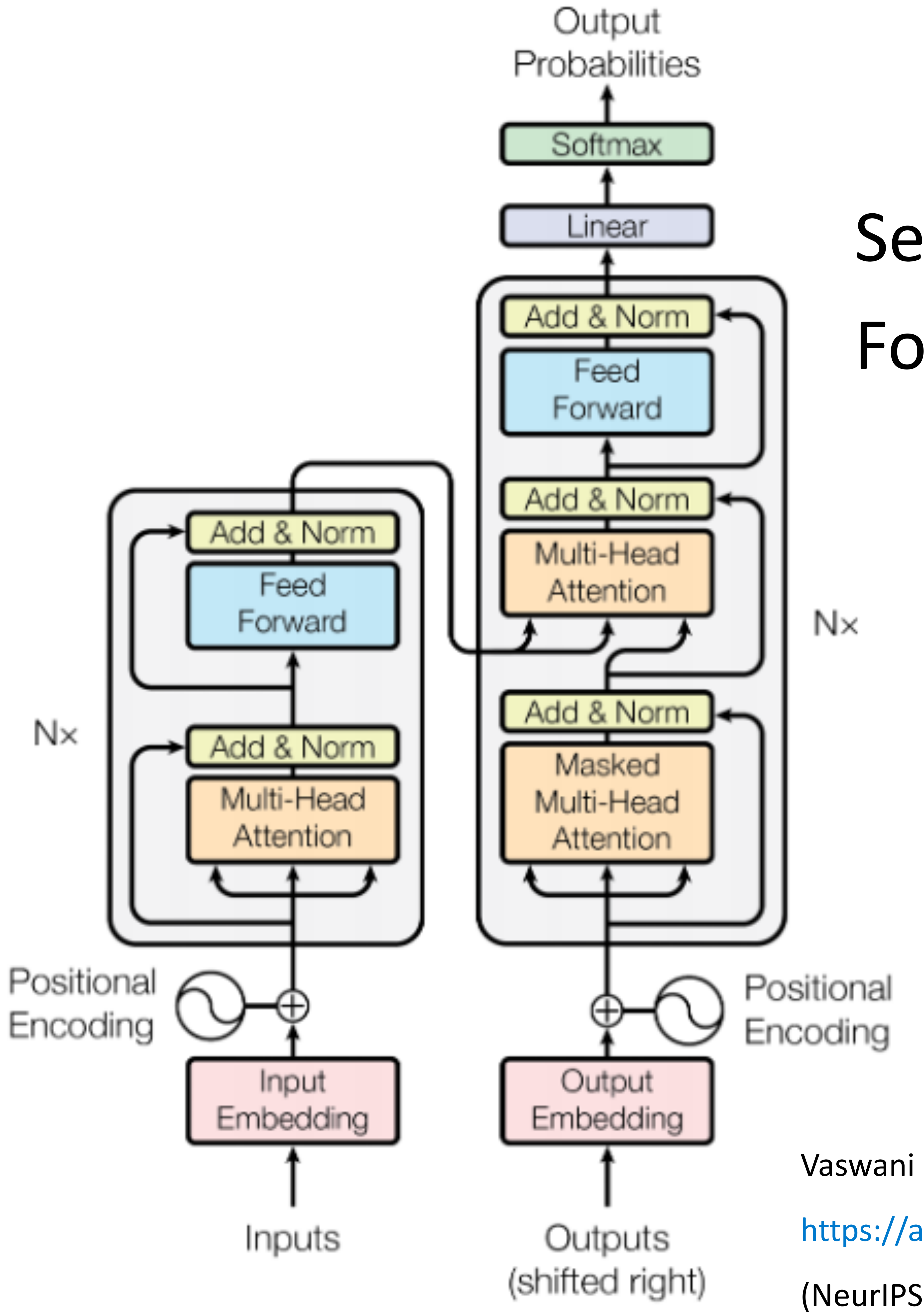
Image credit: Thalles Silva

# Transforming images with GANs

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu Taesung Park Phillip Isola Alexei A. Efros Berkeley AI Research (BAIR) laboratory, UC Berkeley, Nov 2018.

# Novel capacities: Transformers



Sequential data structures

Focus of attention

NLP: correlation in language word sequences

Vision:
Composing remote parts of images

Vaswani *et al*. Attention is all you need]

https://arxiv.org/abs/1706.03762

(NeurIPS 2017)

# Deep learning and NLP

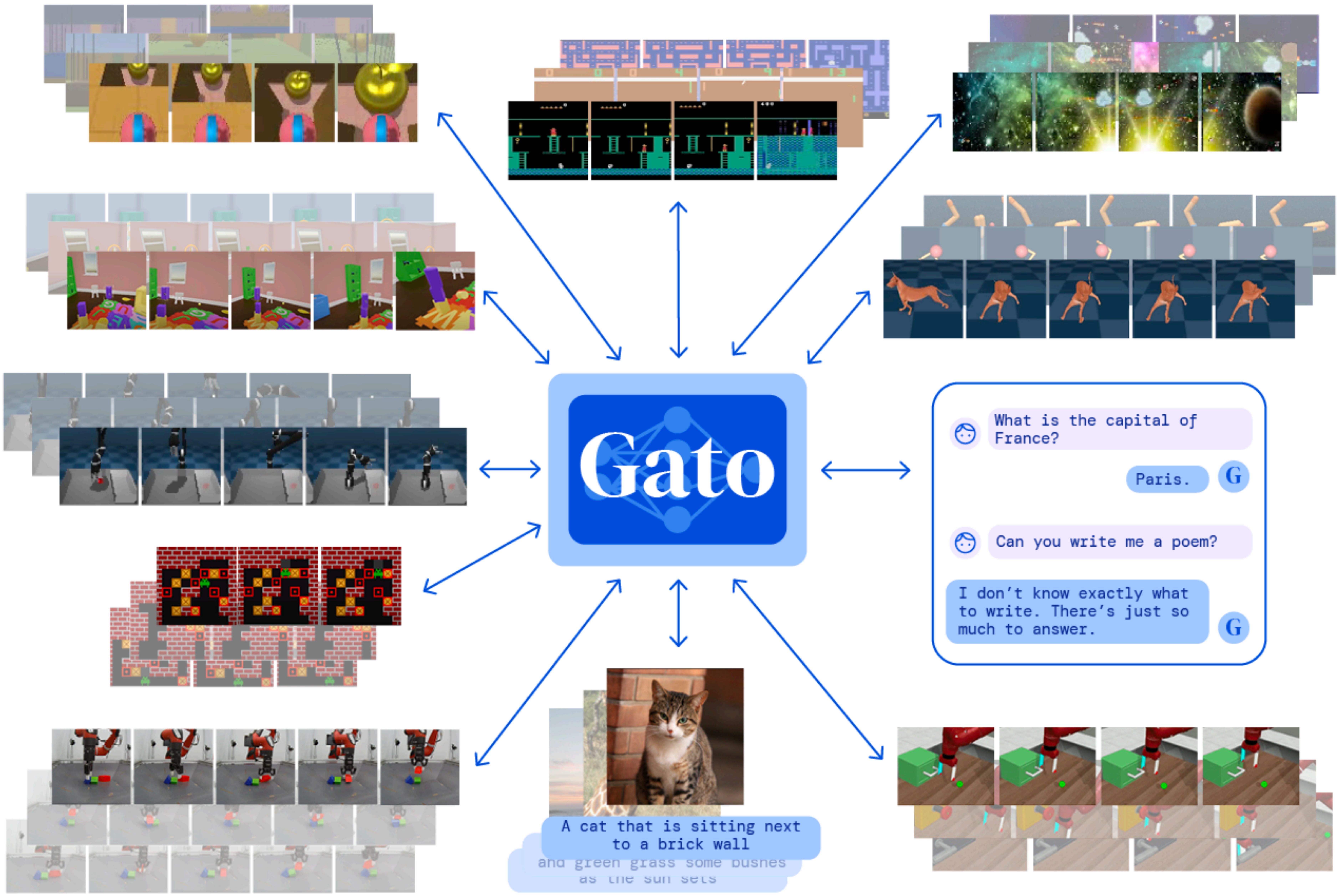
- Large Language models based on unsupervised distribution estimation from a set of examples each composed of variable length sequences of symbols.
- Use Transformers and are re-trained on vast amounts of data from the internet
- GPT 3-4 (OpenAI); BERT; LAMDA (Google); BLOOM (Fr&Hugging Face), …
  GPT3: 175 billion parameters; LAMDA 137B; BLOOM 176 B (59 languages)

Language Models are Unsupervised Multitask Learners.
Alec Radford and Jeffrey Wu and R. Child and David Luan and Dario Amodei and Ilya Sutskever, 2019.

Brown, Tom et al. (31 authors). Language Models are Few-Shot Learners. Neurips 2020

# GATO
# Deepmind, 2022

## A General Purpose System



What is the capital of France?

Paris.

Can you write me a poem?

I don't know exactly what to write. There's just so much to answer.

A cat that is sitting next to a brick wall

and green grass some bushes as the sun sets

# Issues with the Digital transformation (or why do we need Digital Ethics)

- Increasingly imposed use of DT & AI in human activities
- Massive use of (physical and software) automation transforming relation to work, social value of individuals and the economy
- Transformation of human relationships
- Personal data dissemination and privacy breach
- Inherent limitations of digital technologies and AI in particular
- Decisions impacting humans, made by algorithms
- Confusion between human capacities and identity vs. machine “cognitive” capacities, human-like expressions (NL, emotions) , appearance or behaviour
- Impact on the Planet

# Issues with Statistical Machine Learning

- Black box: millions/billions of parameters, optimization algorithms, un certified off-the-shelf components
- No solid verification and validation processes or qualification of results
- Quality and representativeness of data. Data Bias
- Bias due to design and architecture choices
- Inappropriate correlations, absence of causality between data and results
- No explicability
- Computational level: No semantics, no understanding of manipulated symbols, no context awareness
- Environmental cost

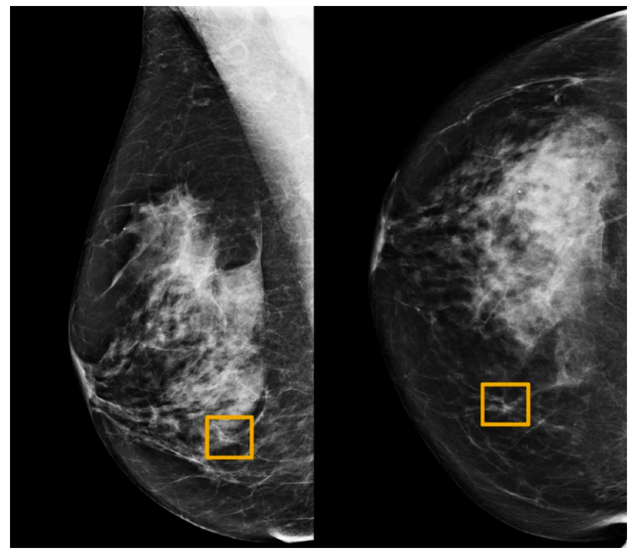
# Risks and Trustworthiness of AI Systems

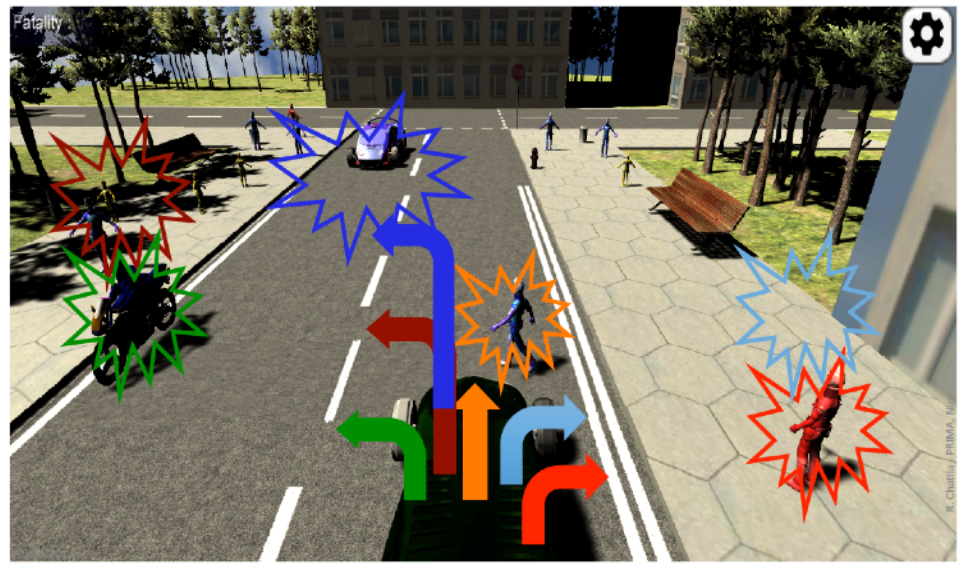
- No ethical rules in academic AI research
- Advanced AI research in industry without ethical oversight
- Applications in critical domains (healthcare, transport, security…)
- Applications potentially threatening human rights and values (surveillance, opinion manipulation, policing, justice, access to jobs and education, …)

→ **Need for robustness and safety**

→ **Need for ethics and governance**

} Transparency
Explainability





# Human Responsibility

- *“To ensure every stakeholder involved in the design and development of AIS is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.”*

  Mission statement of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016

  https://ethicsinaction.ieee.org

*“To support and guide the responsible adoption of AI that is grounded in human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN Sustainable Development Goals.”*

gpai.ai
2020

**Appropriate design approaches and governance frameworks**

# Principles of Biomedical Ethics

- **Respect for autonomy** (the obligation to respect the decision making capacities of autonomous persons);
- **Non-maleficence** (the obligation to avoid causing harm);
- **Beneficence** (obligations to provide benefits),
- **Justice** (obligations of fairness in the distribution of benefits and risks).

Practically, application of these principles mostly translate into the search for a balance between benefits and risks

Beauchamp TL. Methods and principles in biomedical ethics. J Med Ethics. 2003 Oct;29(5):269-74. doi: 10.1136/jme.29.5.269. PMID: 14519835; PMCID: PMC1733784.

Beauchamp TL, Childress J. Principles of biomedical ethics. New York: Oxford University Press, 1st ed, 1979

# Framework for Trustworthy AI (EU HLEG-AI, 2019)

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

AI

ETHICS GUIDELINES
FOR TRUSTWORTHY AI

- **Demonstrable trustworthiness** as a prerequisite to develop, deploy and use AI systems.
  - Trust in organizations and processes for developing and deploying AI
  - Appropriate conditions of use and applications

https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

# Ethical Principles for Trustworthy AI

European Commission

Independent

High-Level Expert Group on Artificial Intelligence

set up by the European Commission

AI

Ethics Guidelines for Trustworthy AI

- **Principle of Autonomy**: “Preserve Human Agency and control”
- **Principle of Non maleficence**: “Do no Harm” - Neither cause nor exacerbate harm  or otherwise adversely affect human beings. Safety and security, technical robustness.
- **Principle of Justice**: “Be Fair”. Equal and just distribution of benefits and costs, free from unfair bias, increase social fairness
- **Principle of <u>Explicability</u>**: “Operate transparently”: Interpretability, traceability, auditability, transparent system capabilities, …

# Key Requirements for Trustworthy AI

## High-Level Expert Group on AI (EU) - April 2019

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES
FOR TRUSTWORTHY AI

1. **Human agency and oversight**- Including respect tof fundamental rights, human control
2. **Technical robustness and safety** - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance** - Including respect for privacy, quality and integrity of data, and access to data
4. **Transparency** - Including traceability, **explainability** and communication
5. **Diversity, non-discrimination and fairness** - Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing** - Including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability** - Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

**Tool: Assessment List for Trustworthy AI - ALTAI**

https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

# Achieving Trustworthy AI : Technical Aspects

- *Architectures for Trustworthy AI*
- *Ethics and rule of law by design (X-by-design)*
- *Robustness, safety, security, testing and validation*
- *Explanation methods*
- *Quality of Service Indicators*

# Achieving Trustworthy AI : Non-Technical Aspects

Independent
High-Level Expert Group on
Artificial Intelligence
set up by the European Commission

Ethics Guidelines
for Trustworthy AI

- *Regulation*
- *Codes of conduct*
- *Standardisation*
- *Certification*
- *Accountability via governance frameworks*
- *Education and awareness to foster an ethical mind-set*
- *Stakeholder participation and social dialogue*
- *Diversity and inclusive design teams*

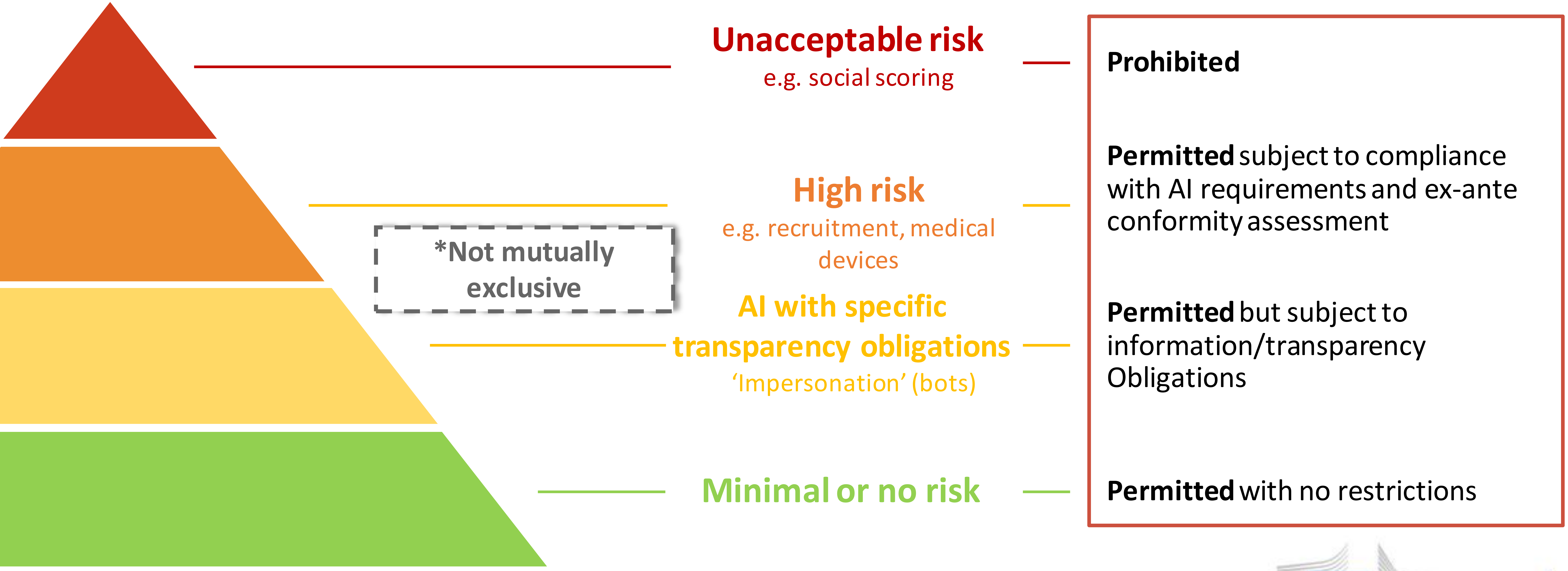
# IEEE P7000™ Standardization Projects for Ethically Aligned Design

IEEE P7000- Model Process for Addressing Ethical Concerns During System Design
IEEE P7001- Transparency of Autonomous System
IEEE P7002- Data Privacy Process
IEEE P7003- Algorithmic Bias Considerations
IEEE P7004- Standard on Child and Student Data Governance
IEEE P7005- Standard on Employer Data Governance
IEEE P7006- Standard on Personal Data AI Agent Working Group
IEEE P7007- Ontological Standard for Ethically driven Robotics and Automation Systems
IEEE P7008- Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
IEEE P7009- Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEE P7010- Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
IEEE P7011- Standard for the Process of Identifying and Rating the Trustworthiness of News Sources
IEEE P7012- Standard for Machine Readable Personal Privacy Terms
IEEE P7013- Inclusion and Application Standards for Automated Facial Analysis Technology
IEEE P7014- Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems

## Ethics Certification Program for Autonomous and Intelligent Systems (CertifAIEd)

# A risk-based approach to regulation

EU Legislative propos



**Unacceptable risk**

e.g. social scoring

**Prohibited**

**High risk**

e.g. recruitment, medical devices

**Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

***Not mutually exclusive**

**AI with specific transparency obligations**

'Impersonation' (bots)

**Permitted** but subject to information/transparency Obligations

**Minimal or no risk**

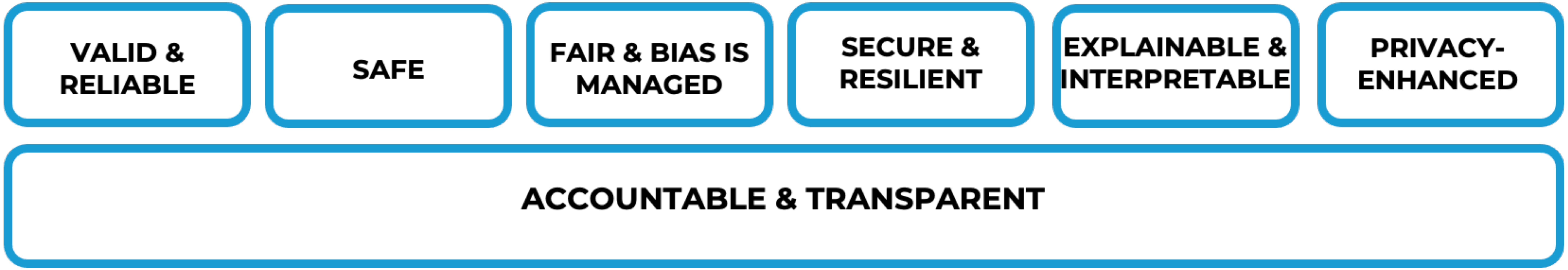
**Permitted** with no restrictions

European Commission

Courtesy L

# NIST’s Trustworthy AI Characteristics

HLEG #2 HLEG #2 HLEG #5 HLEG #2 HLEG #4 HLEG #3



VALID & RELIABLE

SAFE

FAIR & BIAS IS MANAGED

SECURE & RESILIENT

EXPLAINABLE & INTERPRETABLE

PRIVACY-ENHANCED

ACCOUNTABLE & TRANSPARENT

HLEG #4 and 7

NIST AI Risk Management Framework: 2nd Draft
https://www.nist.gov/itl/ai-risk-management-framework

# Ethically Aligned Design
# Value-Based Design
# Value-Sensitive Design

See Standard IEEE P7000-2021 Model Process for Addressing Ethical Concerns During System Design

- Project definition: What are the project objectives? What are its benefits? Does it entail risks? Who are the stakeholders and what are their values*, is AI an adequate tool?
- Project Initiation:  Value conceptualisation; feasibility studies
- System specification: Value analysis, value tensions, value priorities
- Design: Technical solutions to address value priorities, system architecture
- Validation; Success metrics, validation of values and technical solutions
- Deployment and updates
- Evaluation of value compliance during system operation

Examples of values:

physical integrity, physical wellbeing, mental wellbeing, dignity, privacy, freedoms, security, fairness, equality, truth, …

# Takeaways:
# Responsible Development, Use and Governance of AI

- AI is no silver bullet for many application. Avoid technical solutionism.
- ML is a very efficient technology for automating data analysis
- AI systems using machine learning need to be made robust and resilient
- Explainability is essential to build trust in AI systems
- Appropriate design approaches, governance frameworks, auditing and certification of AI systems are necessary.
- Ethical assessment based on the compliance with the HLEG-AI's 7 key requirements, now requested in EU projects.
- Legal framework to in Ethical evaluation of industry projects and applications