# Dependable AI Systems

Homa Alemzadeh
University of Virginia

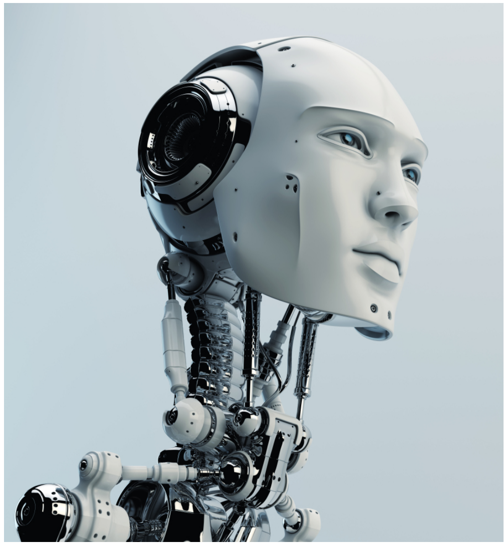
In collaboration with: Kush Varshney, IBM Research

UNIVERSITY of VIRGINIA
ENGINEERING

# Artificial Intelligence

- An intelligent agent or system that perceives its environment and takes actions to maximize chance of success at some goal
- Mimic human cognitive functions

- Central problems or goals of AI:
  - Reasoning
  - Knowledge engineering
  - Planning
  - Learning
  - Natural language processing
  - Perception (vision and speech)



Credits: techcrunch.com

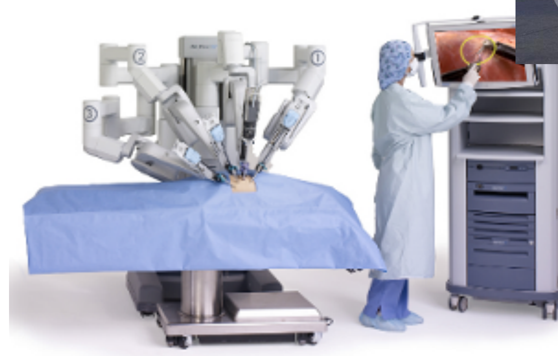
Source: https://en.wikipedia.org/wiki/Artificial_intelligence

# Machine Learning

- Building block of AI systems

- Data Analytics
- Cognitive Systems
- Autonomous Systems
- Cyber-physical Systems



# Basics of Risk Minimization

- **Basic notation:**
- Joint random variables $X \in \mathcal{X}$ (features) and $Y \in \mathcal{Y}$ (labels)
- Probability density function $f_{X,Y}(x, y)$
- A function mapping $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$.
- A loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.
- Risk *R(h)* is defined as the expected value of loss:

$$\mathbb{E}[L(h(X), Y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(h(x), y) f_{X,Y}(x, y) dy dx$$

- L(h(x),Y) measures the discrepancy between the predicted value for y (h(x)) and y itself
- **Ideal Goal:** Learn the function *h(x)* that minimizes the risk *R(h)*.

# ML Empirical Risk Minimization

- In practice, probability distribution of $f_{X,Y}(x,y)$ is unknown
- We only have a training set of samples drawn i.i.d. from the joint distribution (X; Y):

$$\{(x_1, y_1), \ldots, (x_m, y_m)\}$$

- **ML Goal:** Learn the function *h(x)* that such that the empirical risk is minimized:

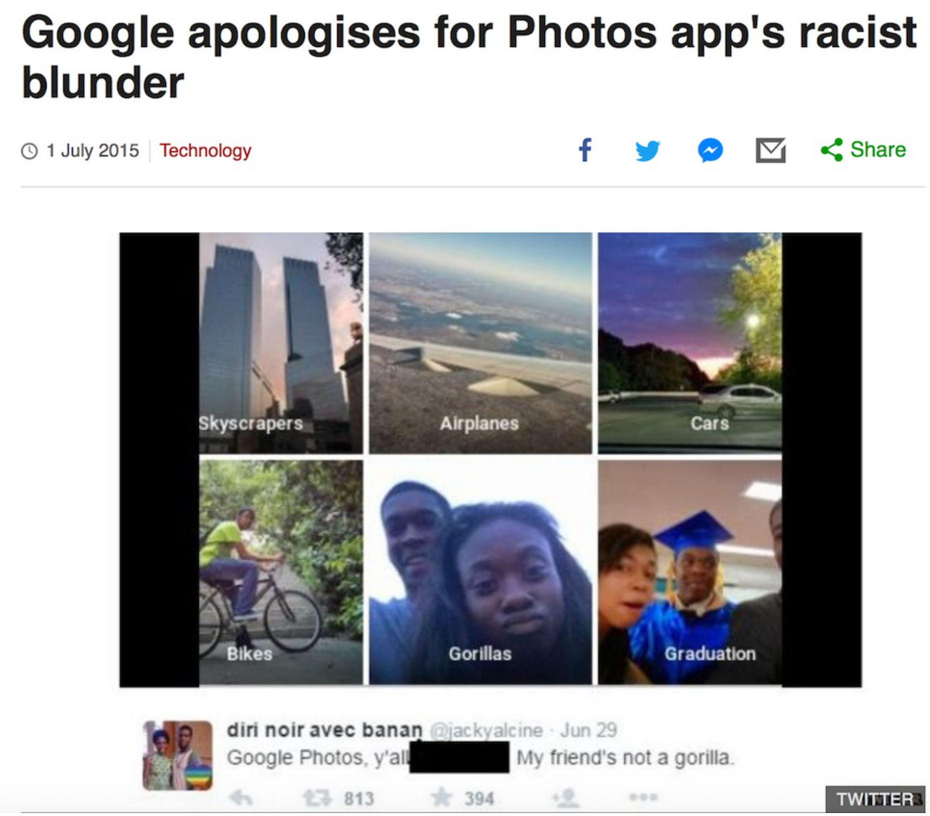
$$R_m^{emp}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i), y_i)$$

# Pitfalls

- Learning systems encounter a finite number of test samples before live deployment
- Actual operational risk is an empirical quantity on the test set
- Training samples (distribution) not always representative of testing samples
- Distribution and cost of outcomes are unknown

# Consequences

[Google Mistakenly Tags Black People as 'Gorillas'](), **2015**



**Google apologises for Photos app's racist blunder**

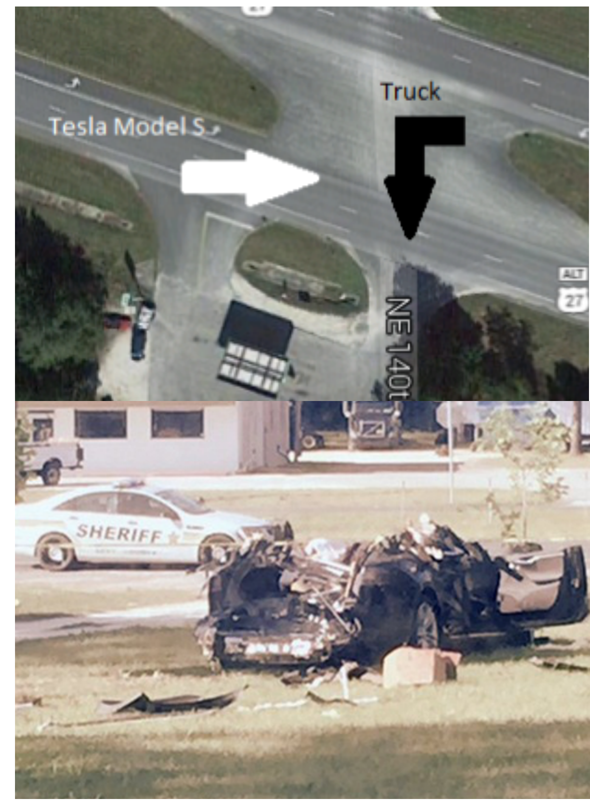
1 July 2015 | Technology

Share



diri noir avec banan @jackyalcine · Jun 29
Google Photos, y'all [redacted] My friend's not a gorilla.
813 394

TWITTER





**Tesla S fatal crash, radar/cameras fail to recognize a white car (2016)**

# Safety of Machine Learning

- Reduction of risk and uncertainty associated with unwanted outcomes that are severe enough to be seen as harmful.
- Both the probability of expected harms and the possibility of unexpected harms.

- **Harmful costs:** Costs of unwanted outcomes must be sufficiently high from society perspective for events to be harmful
- **Epistemic uncertainty:** Harmful outcomes often occur in regimes and operating conditions that are unexpected or undetermined.

- **Safety requirements:**
  - **Consequences:** Harmful to not critical
  - **Costs and impacts:** Real-time, near time, long term

- **Ongoing Research:**
  - Handling Bias in training data
  - Interpretable models

# Security of Machine Learning

- **Evasion Attacks:** find samples that are misclassified by a classifier to evade detection while preserving the desired malicious behavior
- **Poisoning Attacks:** injects constructed samples into the training data to control the properties of the learned model
- **Privacy-Preserving Learning:** collaborative model building without exposing data (multi-party secure computation)
- **Disclosure:** protect sensitive information about the training data from interactions with the model.

Credits: David Evans, Quanquan Gu, Mohammad Mahmoody, Yanjun Qi, CS Department, UVA

# Research Programs

- **Future of Life Institute**
  **AI Safety Research**
  https://futureoflife.org/ai-safety-research/

- AAAI Open Letter: Research priorities for robust and beneficial artificial intelligence
  https://futureoflife.org/data/documents/research_priorities.pdf?x33688


Elon Musk donates $10M to keep AI beneficial

October 12, 2015 / by Max Tegmark

- **National Science Foundation**
- **Intelligent Physical Systems** (IPS)

- **Reflective:** Capable of monitoring their actions, diagnosing problems, and optimizing, reconfiguring, and repairing autonomously.

- **Ethical**: Adhere to an ethical system of societal and legal rules and capable of ethical reasoning, such as incorporating societal values into their reasoning.

**Smart and Autonomous Systems (S&AS)**

**PROGRAM SOLICITATION**
**NSF 16-608**

**National Science Foundation**

Directorate for Computer & Information Science & Engineering
Division of Information & Intelligent Systems
Division of Computer and Network Systems
Division of Computing and Communication Foundations

# Governments Initiatives


- **White House Office of Science and Technology Workshops:**
  - Legal and Governance Implications of Artificial Intelligence
  - Safety and Control for Artificial Intelligence
  - The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term

- **European Union**
  - Regulations for data protection taking effect in 2018
  - Prohibiting algorithms that make any ”decision based solely on automated processing, including profiling” that significantly affect a data subject or produce legal effects concerning him/her.
  - Affecting recommendation systems, credit and insurance risk assessments, and social networks

# Community Activities

- **NIPS** workshop on Reliable Machine Learning in the Wild
  https://sites.google.com/site/wildml2016nips/
- **StartupML** workshop on Adversarial machine learning
  https://conf.startup.ml/adversarial/
- **ISSRE** workshop on Software Certification (WoSoCer 2017)
  **Special theme: Certification of Autonomous/ML/AI-based systems**
  https://sites.google.com/view/wosocer
- **DSN workshop on Dependable ML/AI Systems**

# References


- Kush R. Varshney and Homa Alemzadeh. On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products. CoRR, abs/1610.01256, 2016.
- Weilin Xu, et al. “Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers,” In Proceedings of the Network and Distributed Systems Symposium , 2016.
- Battista Biggio, et al, “Poisoning Attacks against Support Vector Machines,” In Proceedings of the 29th International Conference on Machine Learning , 2012.
- Lu Tian, et al. “Aggregating Private Sparse Learning Models Using,” Multi-Party Computation. In Private MultiParty Machine Learning (NIPS 2016 Workshop) , December 2016.
- Preparing for the Future of Artificial Intelligence, https://obamawhitehouse.archives.gov/blog/2016/05/03/preparing-future-artificial-intelligence
- B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a ’right to explanation’,” in Proc. ICML Workshop Human Interpretability, New York, NY, Jun. 2016, pp. 26– 30.