UNIVERSITÉ DE GENÈVE

Side Effects of AI Revolution

Slava Voloshynovskiy

University of Geneva

Switzerland

2019, Geneva

SIP stochastic information processing

Outline

- Modern AI: beliefs and reality
- Open issues:
  - AI and IP issues
  - AI privacy and impact on humans
  - AI discrimination
  - AI vs AI: AI vulnerability and responsibility
  - AI vs humans: DeepFakes
  - AI and ethical issues in autonomous vehicles
- Main lessons and conclusions

Great advancement of AI technologies in various spheres

- **Health and medical care**

  Monitoring, diagnostics, treatment

- **Finance**

  Investment, trading, banking

- **Security**

  Private: biometrics

  Public: surveillance, automatic recognition, tracking, etc.

- **Science and discovery**

  Genetics, biology, astrophysics, high energy physics, etc.

- **Technology**

  Smart manufacturing, transportation, cities, connected objects

The main factors of AI success:

- **Big data**: availability of massive training datasets
- **Hardware**: high performance graphical cards and parallel computing
- **Software**: improved algorithms and new models

The powerful stimulating factors:

- IT giants propose their facilities for computing, storage and AI services
- New AI tendencies in industry, security and science
- High openness of society to accept AI technologies

What is AI, machine learning and deep learning?

Artificial Intelligence

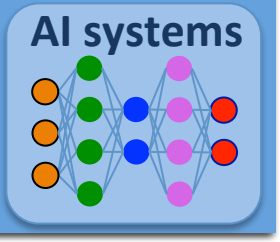
Any technique that enables computers to mimic human behavior

Machine Learning

Ability to learn without explicitly being programmed

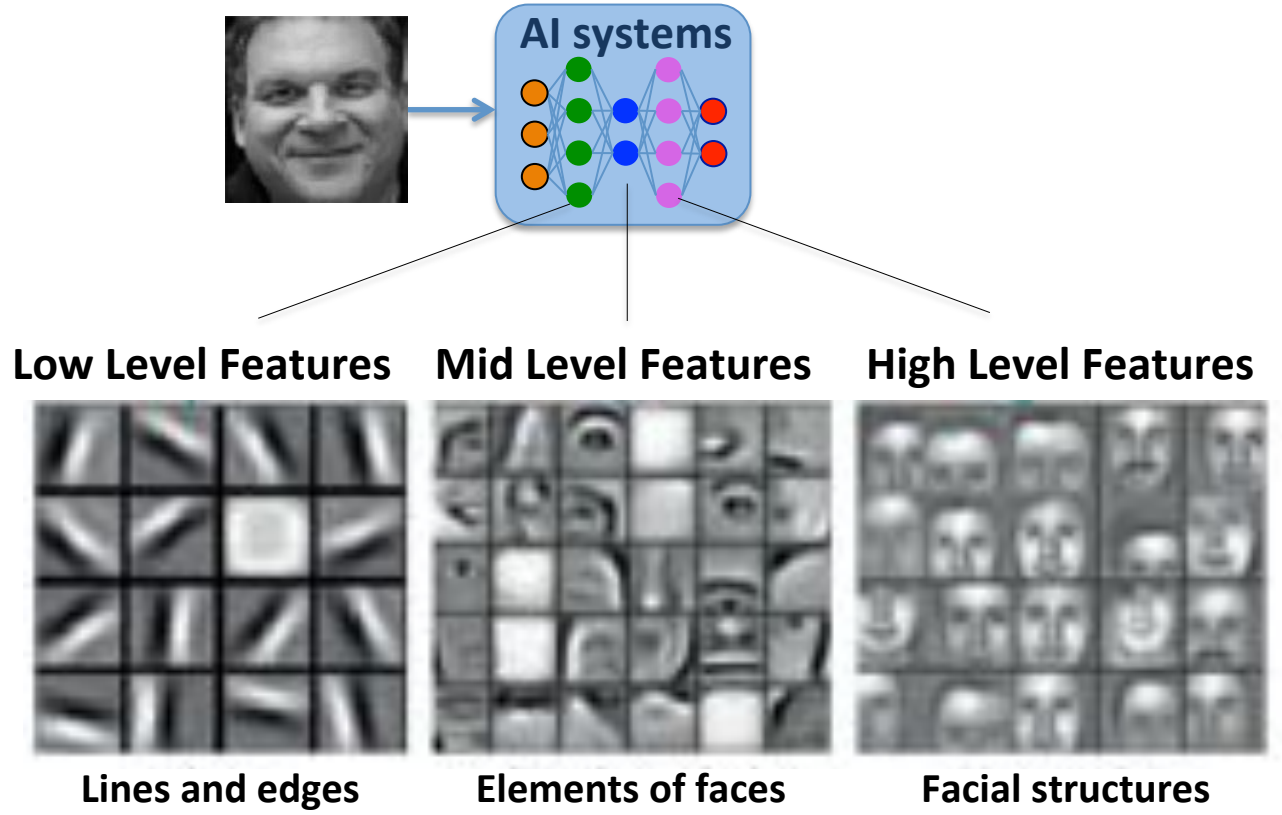
Deep Learning

Learn underlying features in data using deep neural network



What is special about deep learning?

- **Hand engineered features:** time consuming and not scalable
- **Deep learning**: a direct learning of underlying features from data with some semantic meaning



Common beliefs about AI

- AI is just about robots
- AI is fully automated and does not need any human intervention
- AI is very similar to the human brain
- AI is created by definition to be ethical and can not harm or can not be used for harm
- AI algorithms are fully understood and the results produced by them can be trusted due to the open and transparent research
- AI is very close to people in their ability to solve very complex problems

AI reality is that today:

- We have only engineered AI to solve very narrow and dedicated problems like image or speech recognition, image generation, etc.
- However, the **machine reasoning** is still not achieved
- Being faster and more productive in computations, it does not mean that AI is smarter
- AI can be considered as a "**glorified signal processing tool**"

  …. yet with many weaknesses, open technical, legal and ethical issues.

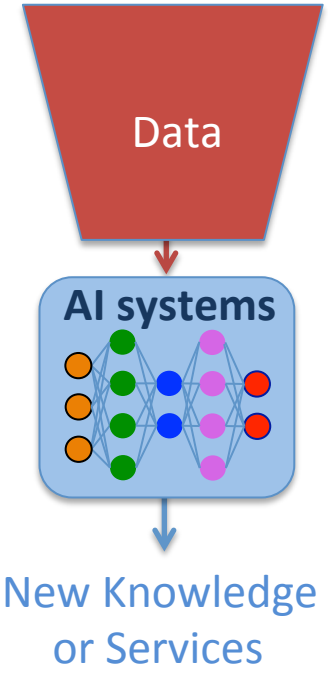
Goal of this presentation:

by demonstrating a great advancement of AI technologies to highlight major classes of open issues requiring interdisciplinary discussion

Common belief about AI training

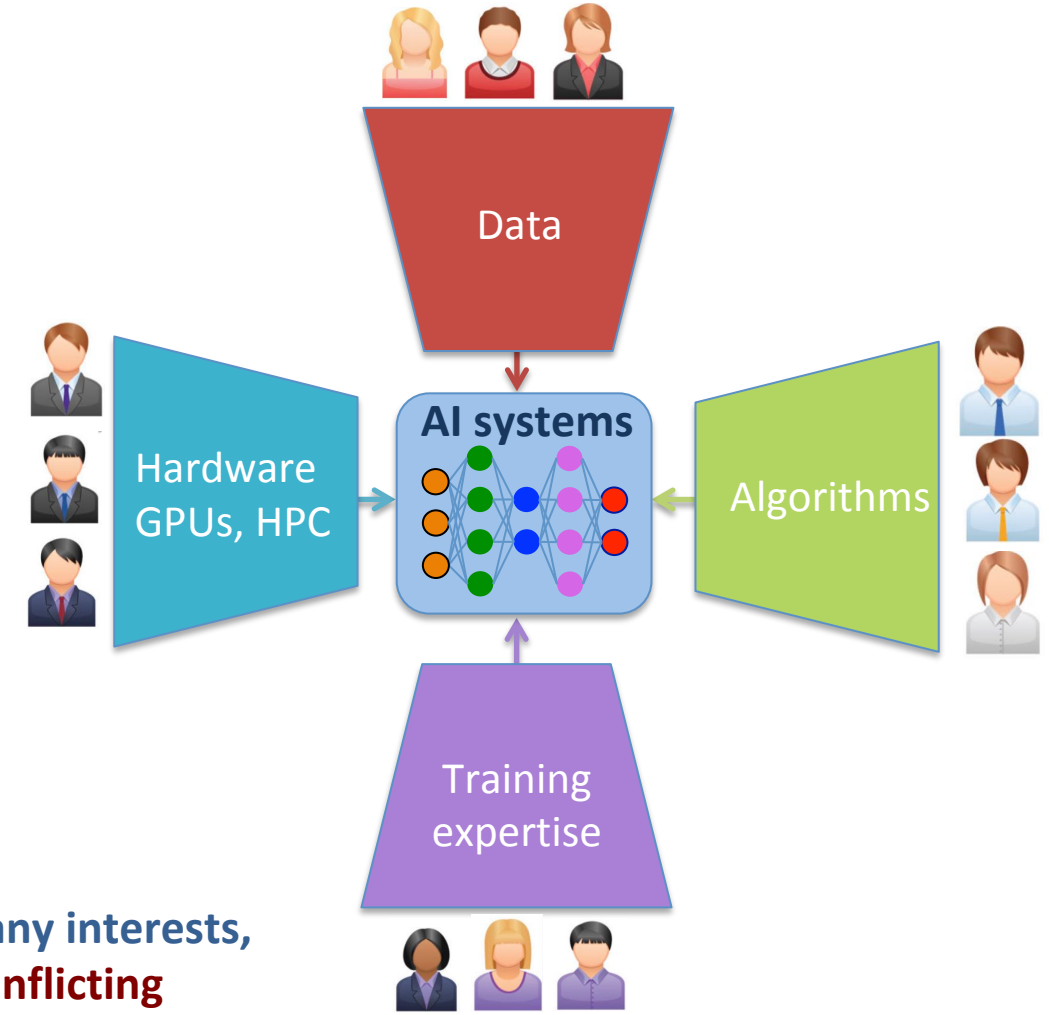
Belief

AI training is fully automated



Reality

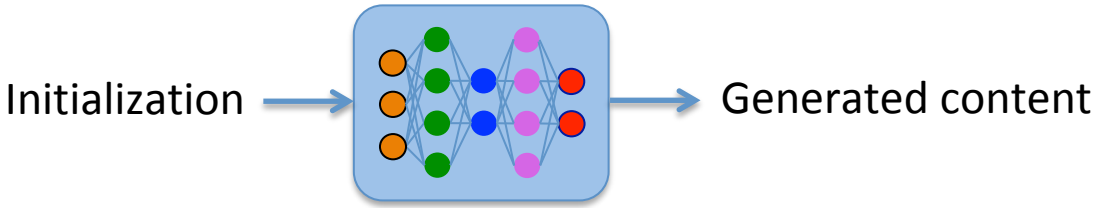
massive engagement of people



Many people = many interests,
sometimes conflicting

IP issues with respect to the results produced by AI

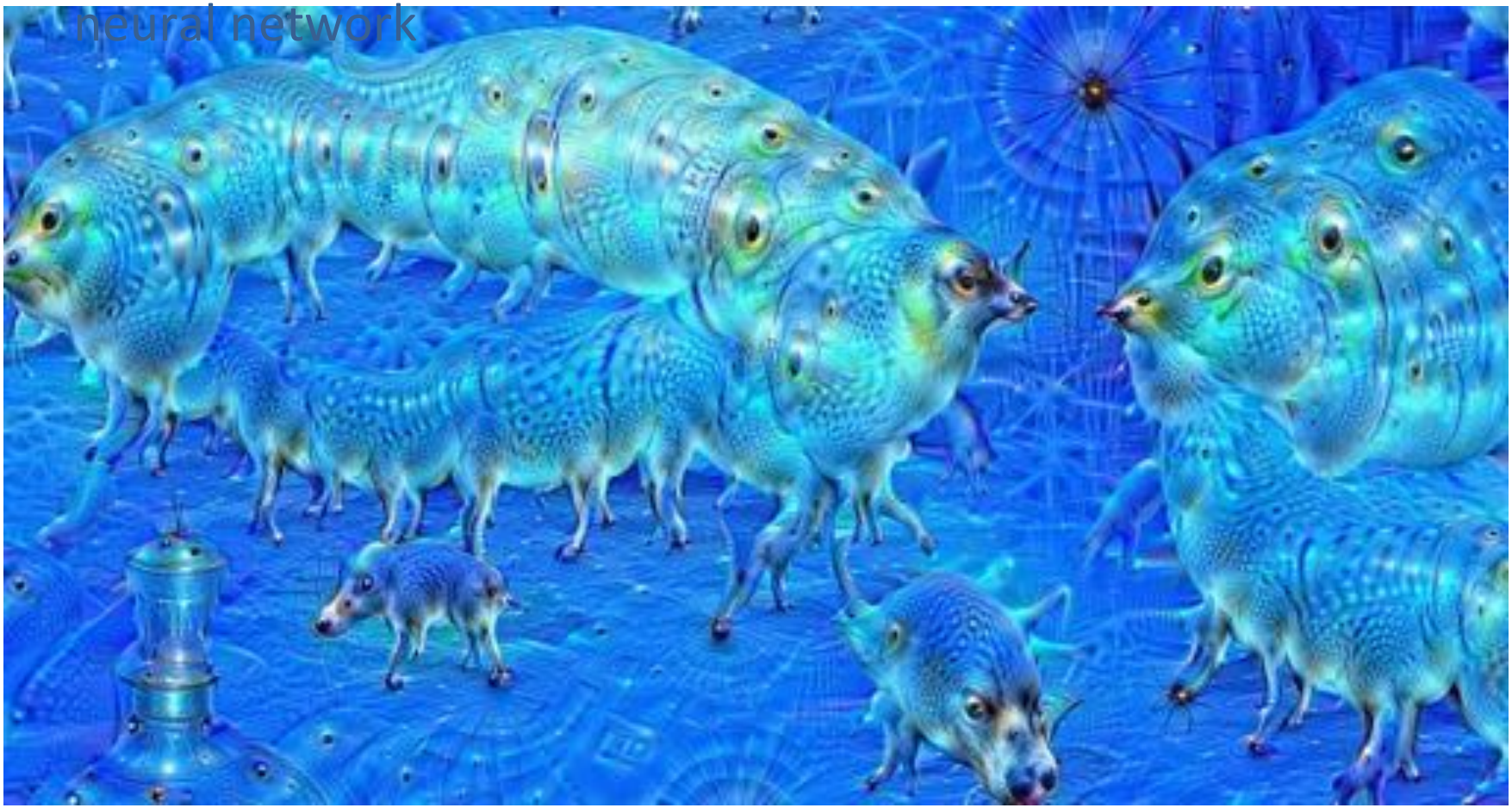
- Given a trained AI system
- By properly initializing, this AI can generate a new content such as art, music, design, etc.



https://link.estadao.com.br/noticias/inovacao,veja-as-imagens-criadas-pelo-sistema-de-redes-neurais-do-google,10000029179

IP issues with respect to the results produced by AI

A strange visual picture created by the Google Dream neural network



https://medium.freecodecamp.org/how-to-create-a-neural-network-in-javascript-in-only-30-lines-of-code-343dafc50d49

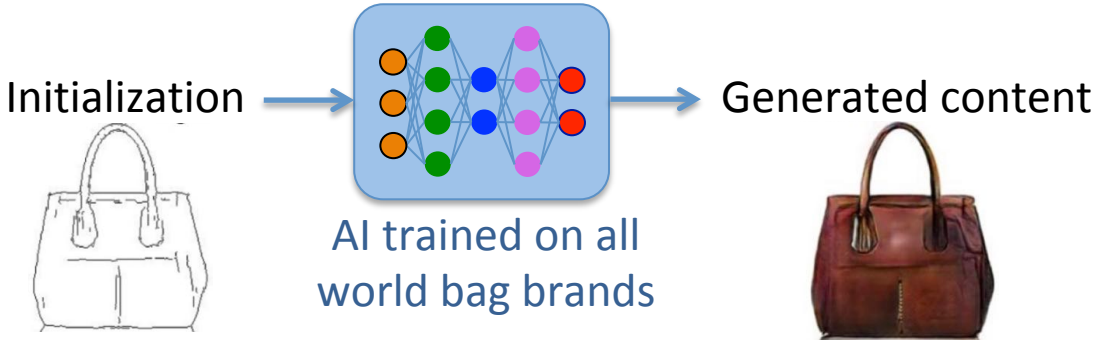
IP issues with respect to the results produced by AI

Who is an owner of newly generated content?

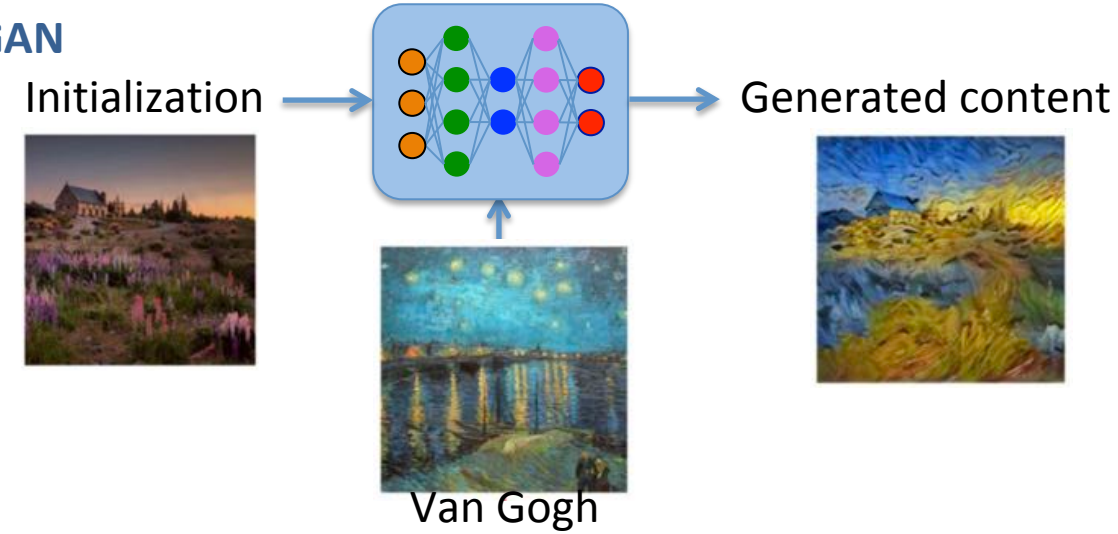
https://link.estadao.com.br/noticias/inovacao,veja-as-imagens-criadas-pelo-sistema-de-redes-neurais-do-google,10000029179

IP issues with respect to the results produced by AI

- **IP issue**: who is an owner of newly generated content?

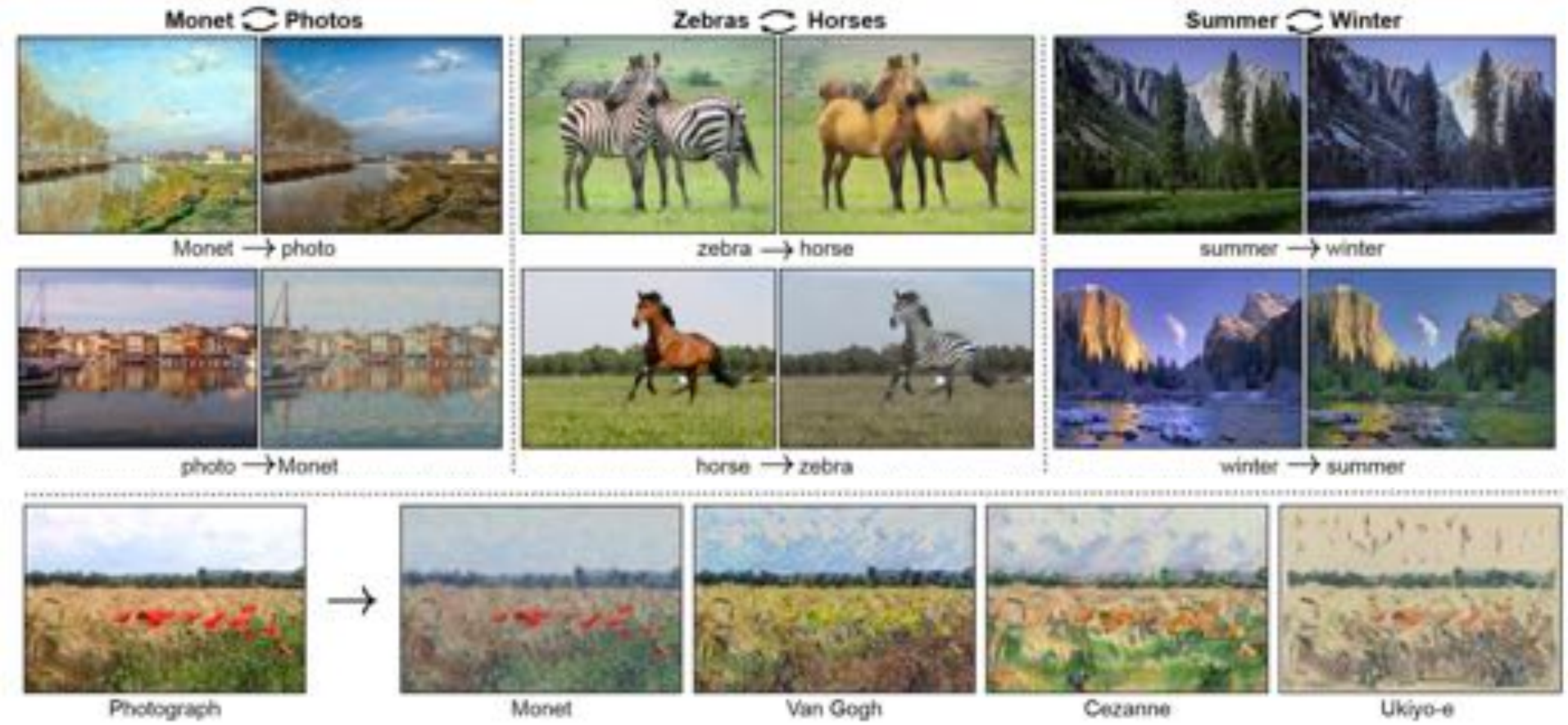


CycleGAN



IP issues with respect to the results produced by AI

- **IP issue**: who is an owner of newly generated content?



J.-Y. Zhu *et. al.*, Unpaired Image-to-image translation using Cycle-consistent adversarial networks, **https://junyanz.github.io/CycleGAN/**

IP issues with respect to the results produced by AI

- **Point of view A:**
  - The **owners** are people **who provided the trained AI system**: data owner(s), algorithm developer, AI training engineer, etc.
  - The person who has initiated the system and selected the content does not contribute much to this process

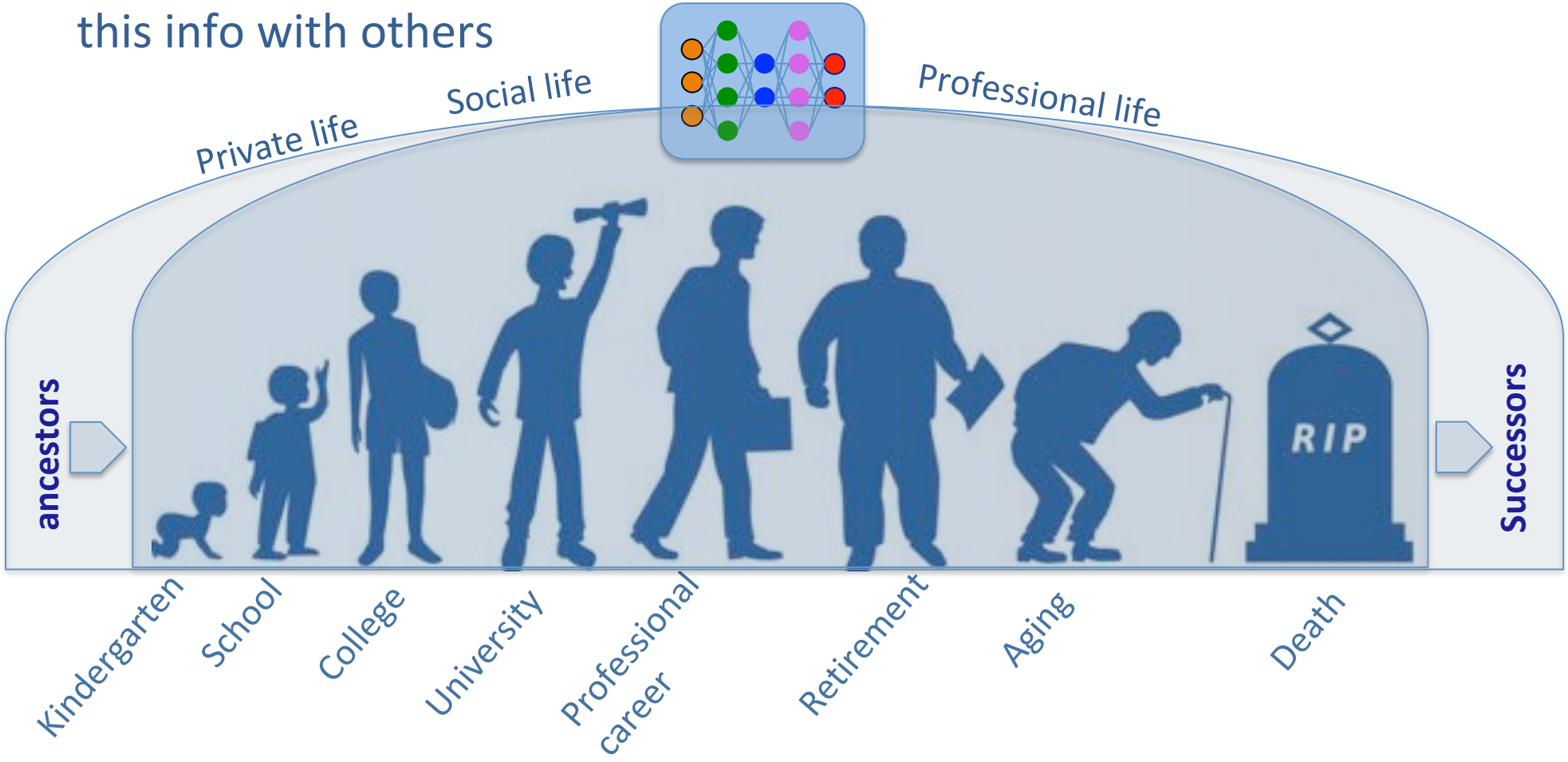
- **Point of view B:**
  - The **owner** is a person **who initiated the generation process** and chosen the most interesting outcome of this process
  - All other players are just technology providers

    Example: an artists uses canvas, inks, brushes, etc. but the providers of these technical means do not pretend on the future IP

AI privacy and impact on humans

AI covers an entire life cycle of humans:

collects data, evaluates, recommends, encourages or discourages, measuring performance, motivates, etc. and potentially AI shares this info with others



AI impact (not always positive) on humans:

- **Performance evaluation systems** starting from school – directly impacting humans
- **Hiring systems** based on AI providing more selective procedures that can be potentially biased – might determine future of people
- **Job markets**: AI will offer better and faster services in many fields thus replacing millions of humans yet avoiding a competition with humans by recommending various options that maybe be again biased

- **Discrimination** due to the unbalanced training data
- **Domination** of some corporations and countries having access to the massive training data; better services; higher profit

We are “under” the influence of AI systems

- **AI systems, are they so perfect that we can entirely trust them our lives and future of our children?**

We consider several examples of **AI imperfection**:

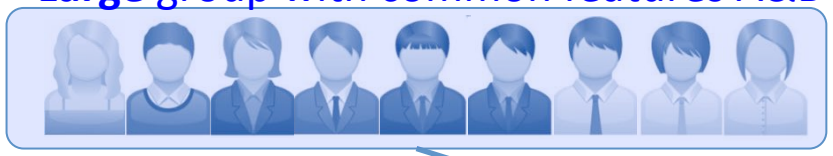
- AI bias and discrimination
- AI vulnerability to adversarial attacks

and associated problems

Bias in learning and risk of "AI discrimination"

Misbalance in training data for different groups

Large group with common features A&B



Small group with common features C&D

AI System Training

Large misbalance in representation of groups

Person from **large** group

AI system

High accuracy of service:
disease prediction, recommendation, taxation, recruitment, etc.

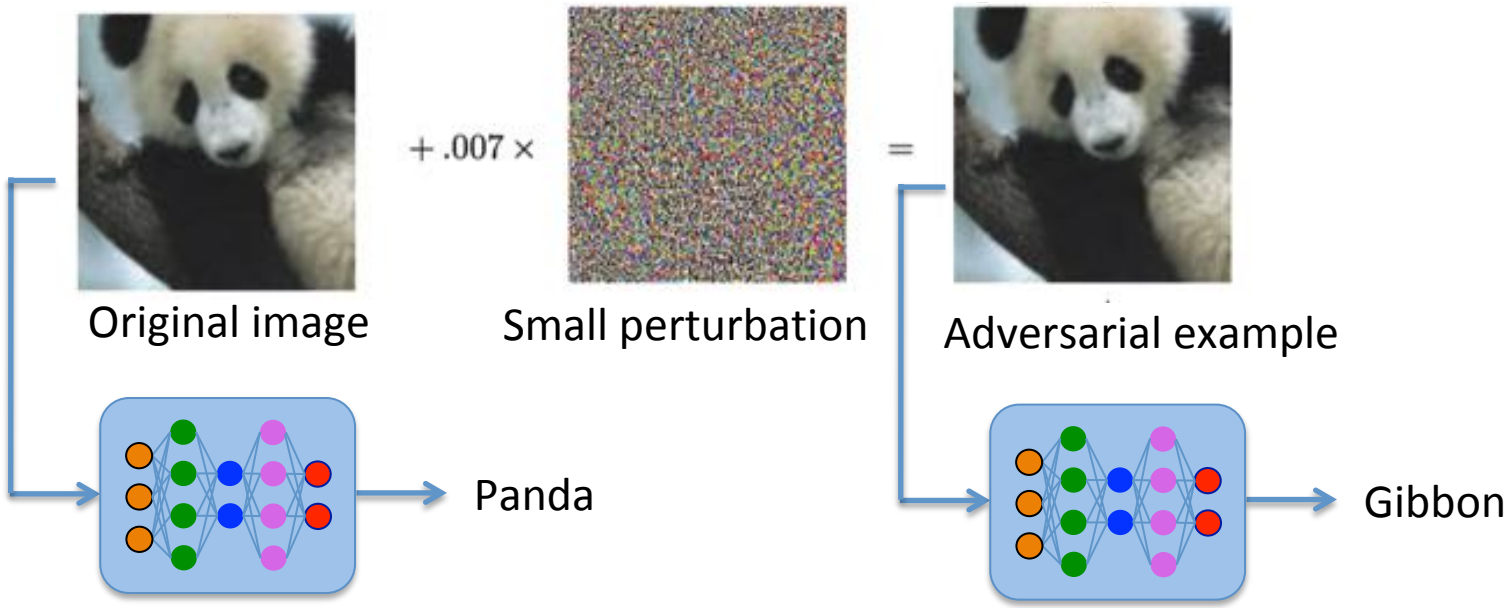
Person from **small** group

AI system

Inaccurate service
Misjudgment
Potential discrimination

Vulnerability of AI in face of adversarial examples

- Given a trained AI system
- Present a specially design **adversarial example** to impact the performance of AI system

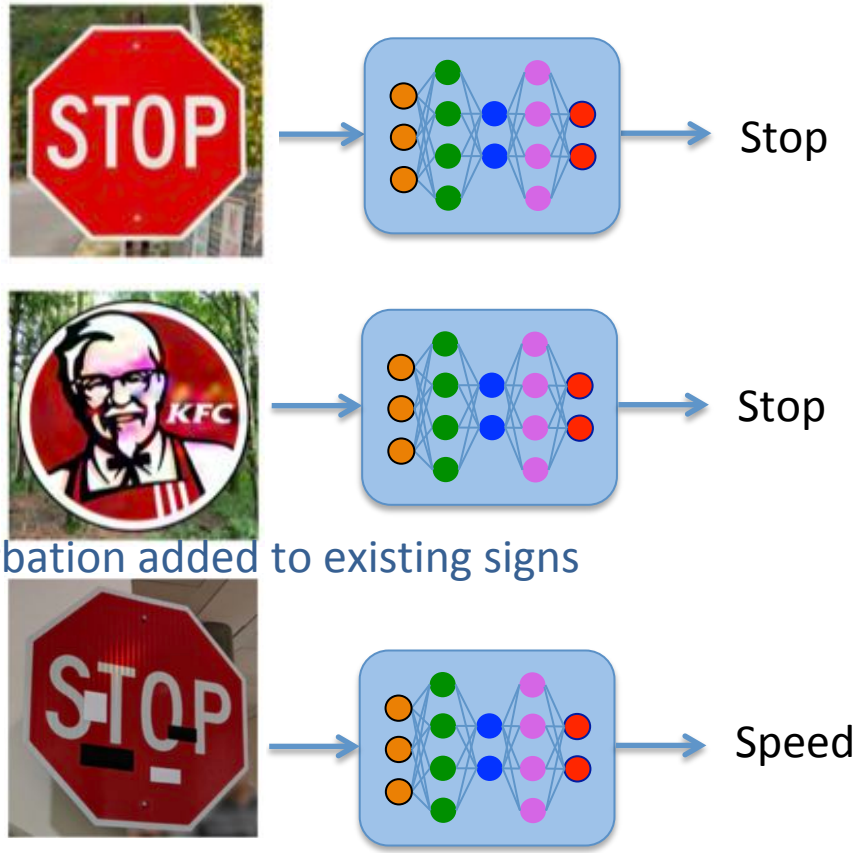


I. Goodfellow et. al, Explaining and harnessing adversarial examples**, ICLR'15**

Vulnerability of AI in face of adversarial examples

- Given a trained AI system
- Physical world attacks: misclassification of road signs

Specially produced signs



Perturbation added to existing signs

Sample video during drive-by tests

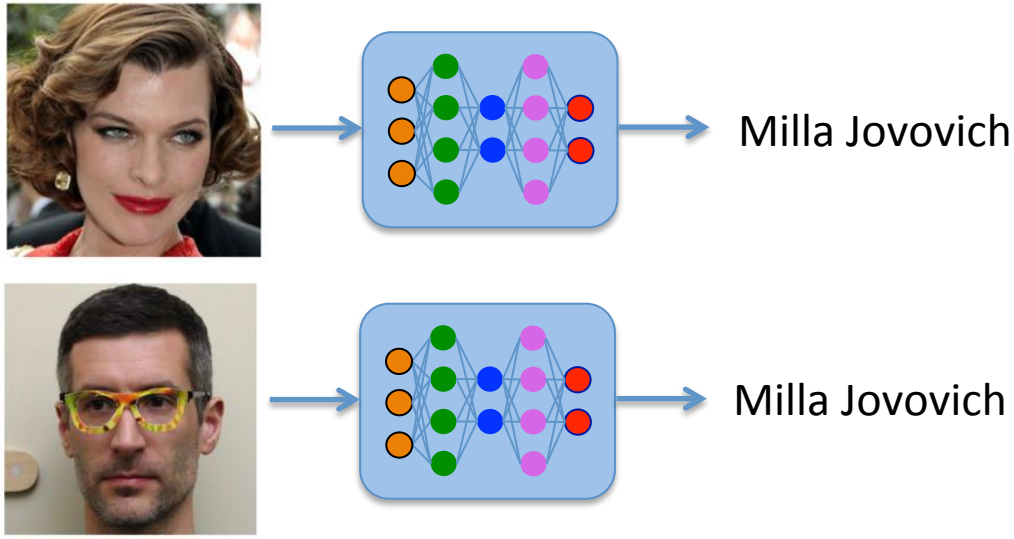
Logo attack classified as ”No passing’ sign with a confidence of 1.0

C. Sitawarin et. al., Rogue Signs, **https://arxiv.org/pdf/1801.02780.pdf**

K. Eykholt et. al., Robust Physical-World Attacks, **https://arxiv.org/pdf/1707.08945.pdf**

Vulnerability of AI in face of adversarial examples

- Given a trained AI system
- Physical world attacks: impersonating people by wearing special accessories in a form of classes



M. Sharif et. Al, Accessorize to a Crime, **ACM CCS'16**

Vulnerability of AI in face of Adversarial Attacks

Conclusion:

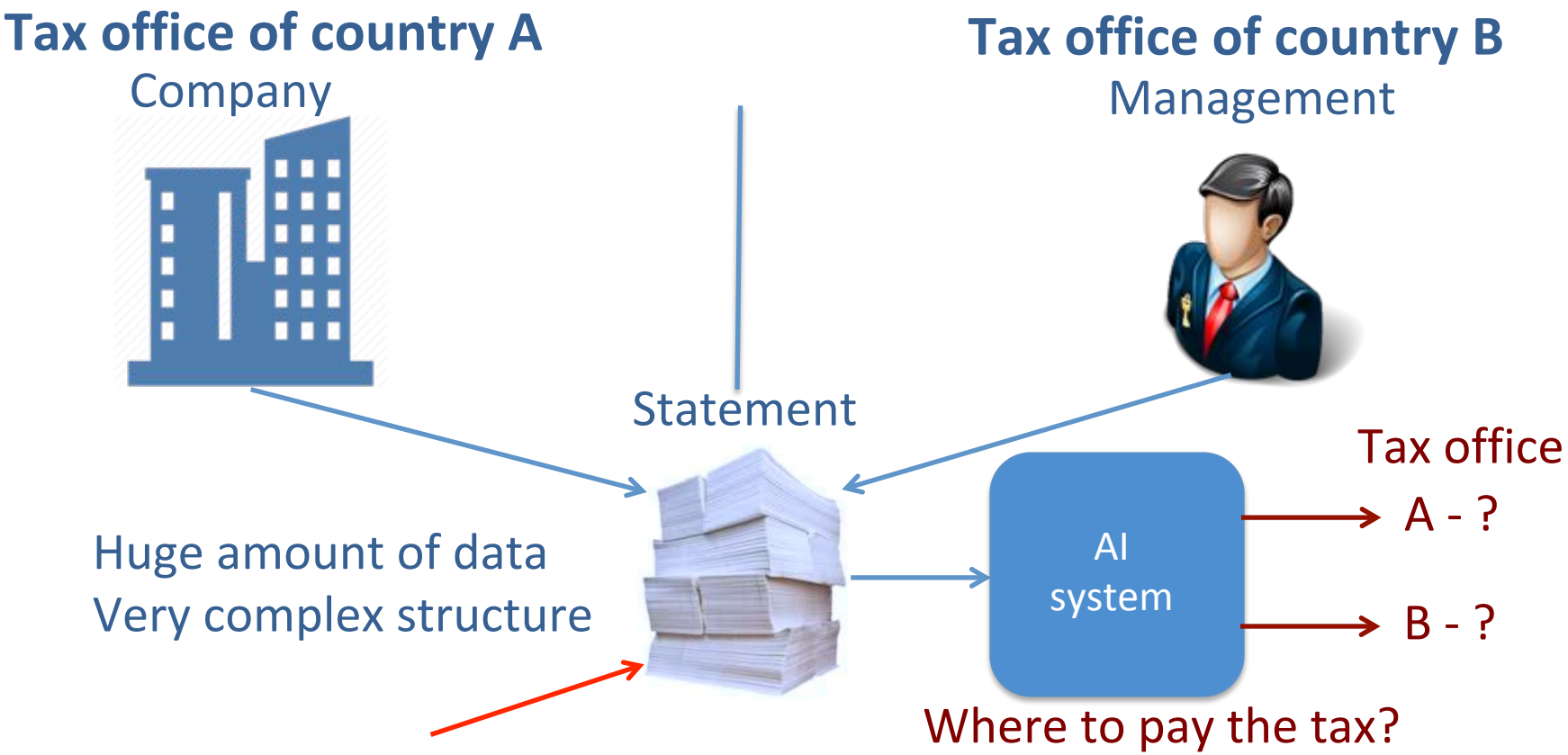
- **Many AI systems are vulnerable to adversarial attacks**
- **Mechanism of defense against these attacks is not well understood**

Question:

- Are we ready to use fragile AI systems in critical applications?
- Likely “No”

… and what about **finance** and **responsability for the wrong AI decisions**?

Vulnerability to adversarial attacks in financial docs

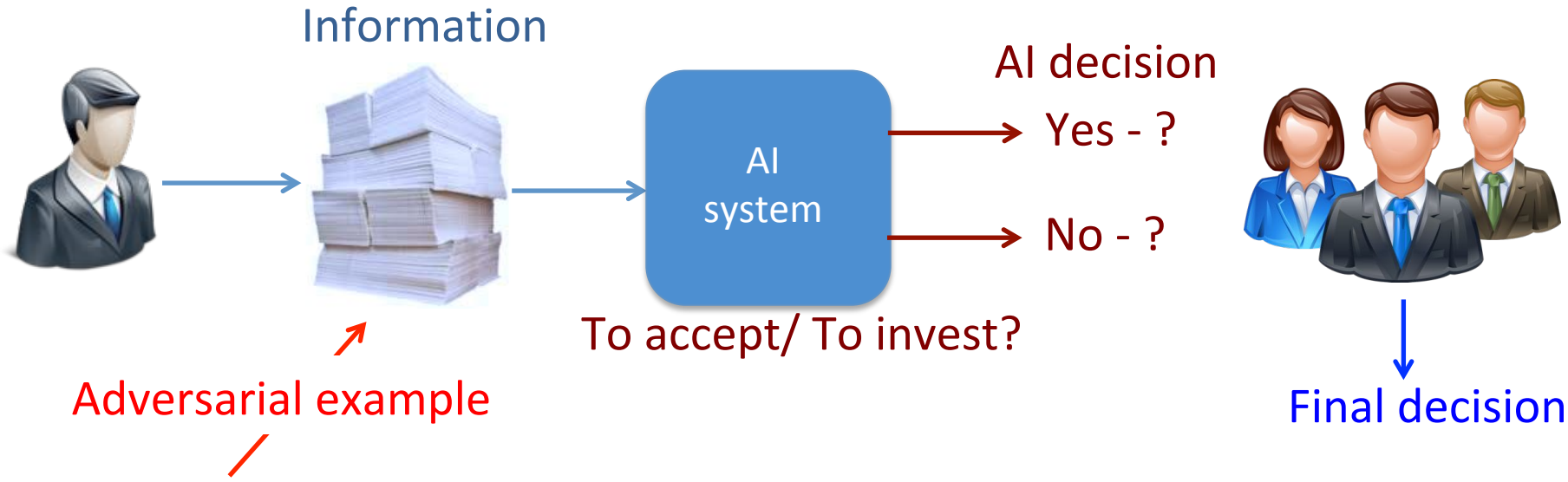


Issue of vulnerability to "attacks":

Knowing how the AI is functioning, i.e., a decision boundary, one can modify (slightly) data in such a way to achieve the most favorable decision for the tax payer

Responsibility for complex/hybrid decisions

- Responsibility for classifying/accepting new clients
- Responsibility for placement or investment



Issue of responsibility for mistakes or vulnerability of AI system:

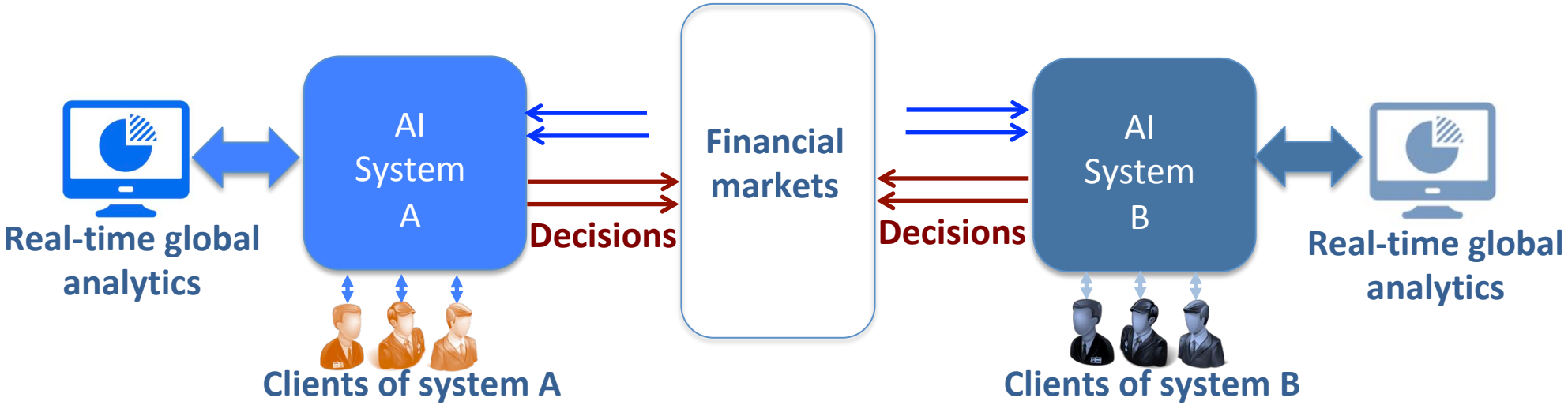
Who will be responsible for the final mistake:

- AI developers or AI service providers?
- people taking the final decision based on a preliminary AI decision?

Responsibility for complex/hybrid decisions

Financial markets:

- AI systems will replace humans due to the need to take fast decisions in face of huge amount of data
- AI systems vs AI systems



Issues of responsibility and control for the decisions and actions of AI systems matching markets and clients' profiles

AI vs humans: DeepFakes

- AI can create synthetic images, videos and audios that are undistinguishable from realistic ones a.k.a. **DeepFakes**
- DeepFakes can be used to fool people or biometric systems

AI vs humans: DeepFakes

Societal and political issues related to consequences of such kind of fakes.

Ethical issues in private and public security

Autonomous vehicles: a great advancement of

- Self-driving cars
- Drones

Uber self-driving car killed a pedestrian on March 19, 2018

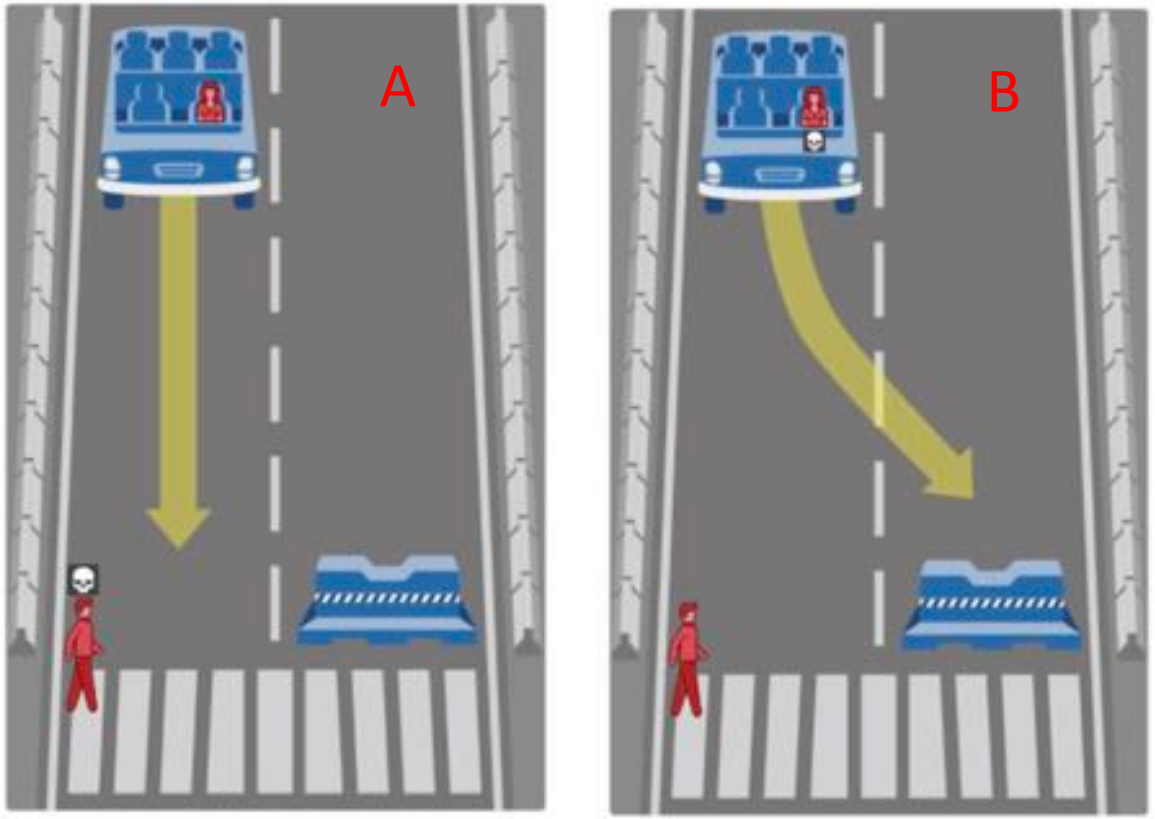
Ethical and **legal** issues around:

- Could a human driver have avoided this crash?
- Even if this crash was avoidable, are self-driving cars still generally safer than human-driven cars?

Ethical issues versus private and public security

- Imagine a situation:

  self driving car in face of dilemma to kill a pedestrian or unavoidable crash of car



What should the self-driving car do? http://moralmachine.mit.edu/

Will you buy such a car knowing that your life is of secondary priority?

AI misusage

- **Open AI:** to grow the trust into AI, researchers and funding agencies encouraging to produce open AI with public code and reproducible results
- **Performance:** these algorithms solve very complex tasks in computer vision, pattern recognition, tracking, etc.
- **Peaceful usage:** when we release these AI algorithms we do not assume that some companies or countries can misuse them in some military applications to create weapon driven by AI
- **Governmental control:** the distribution of weapon is under the strict governmental and international control

AI case: unfortunately, it is virtually impossible to check it out where and how an algorithm from the public domain is used in some military or proprietary application

Main lessons and conclusions

- **AI is a great tool that might revolutionize many fields**
- **The main question is: are we ready for this revolution in view of**
  - Vulnerability of AI systems to adversarial attacks
  - Weak understanding of all factors defining the functioning of complex AI systems
  - Privacy issues
  - Legal questions covering:
    - IP issues
    - Responsibility for the actions and decisions
  - Ethical and misusage issues

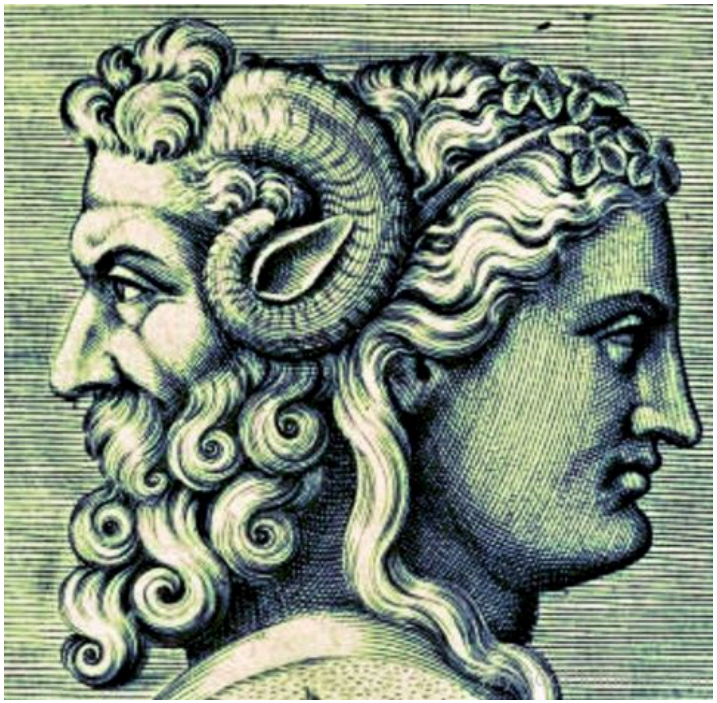
This recalls a need of close interdisciplinary cooperation to be prepared for all “side effects” of AI revolution

Our wish and efforts

To make AI with a «human face»

Currently

AI has a «double face»



It is our responsibility to shape AI for future generations

https://www.deviantart.com/techgnotic/journal/The-Fantastic-Voyage-Of-Isaac-Asimov-503769835

https://gnosticwarrior.com/january-god-janus.html