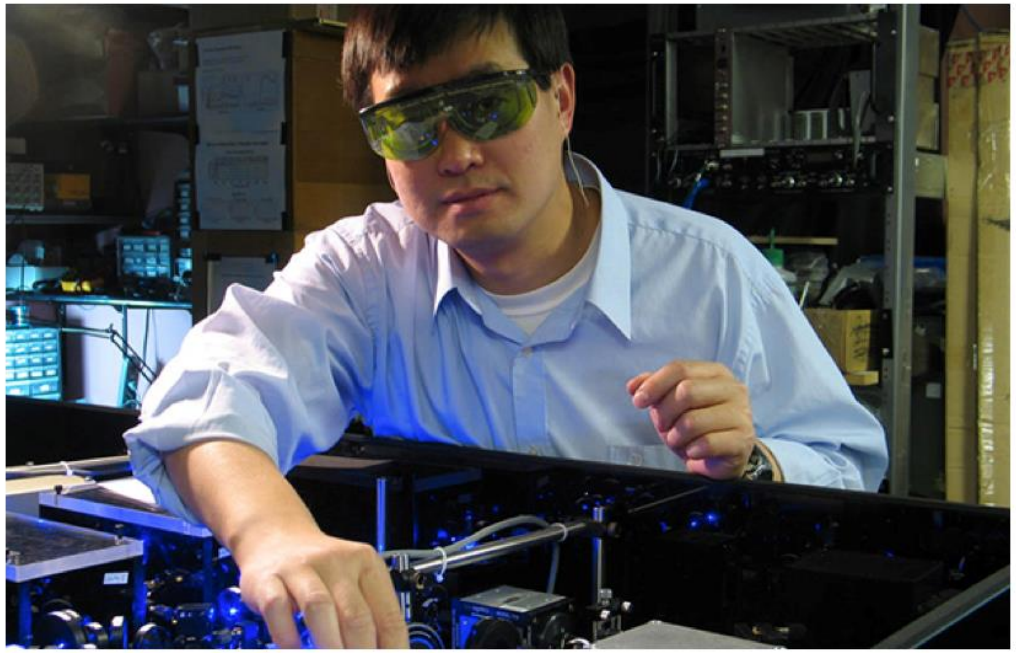
AI
Friend or Foe?

Timothy Grance

Cultivating Trust in IT and Metrology



National Institute of Standards and Technology
U.S. Department of Commerce

ITL's work

Fundamental Research

Adoption

Image Credit: wsj.com

Fundamental and Applied Research and Standards for AI Technologies (FARSAIT)

Applied

Catalyzing the use of machine learning and AI within NIST scientific program.

Fundamental

Measure and enhance the security and trustworthiness of AI systems.

Trustworthy AI

Trustworthy AI

accuracy

reliability

privacy

robustness

explainability

Secure AI

Terminology and Taxonomy of attacks and defenses for Adversarial Machine Learning.

NCCoE
NATIONAL CYBERSECURITY CENTER OF EXCELLENCE

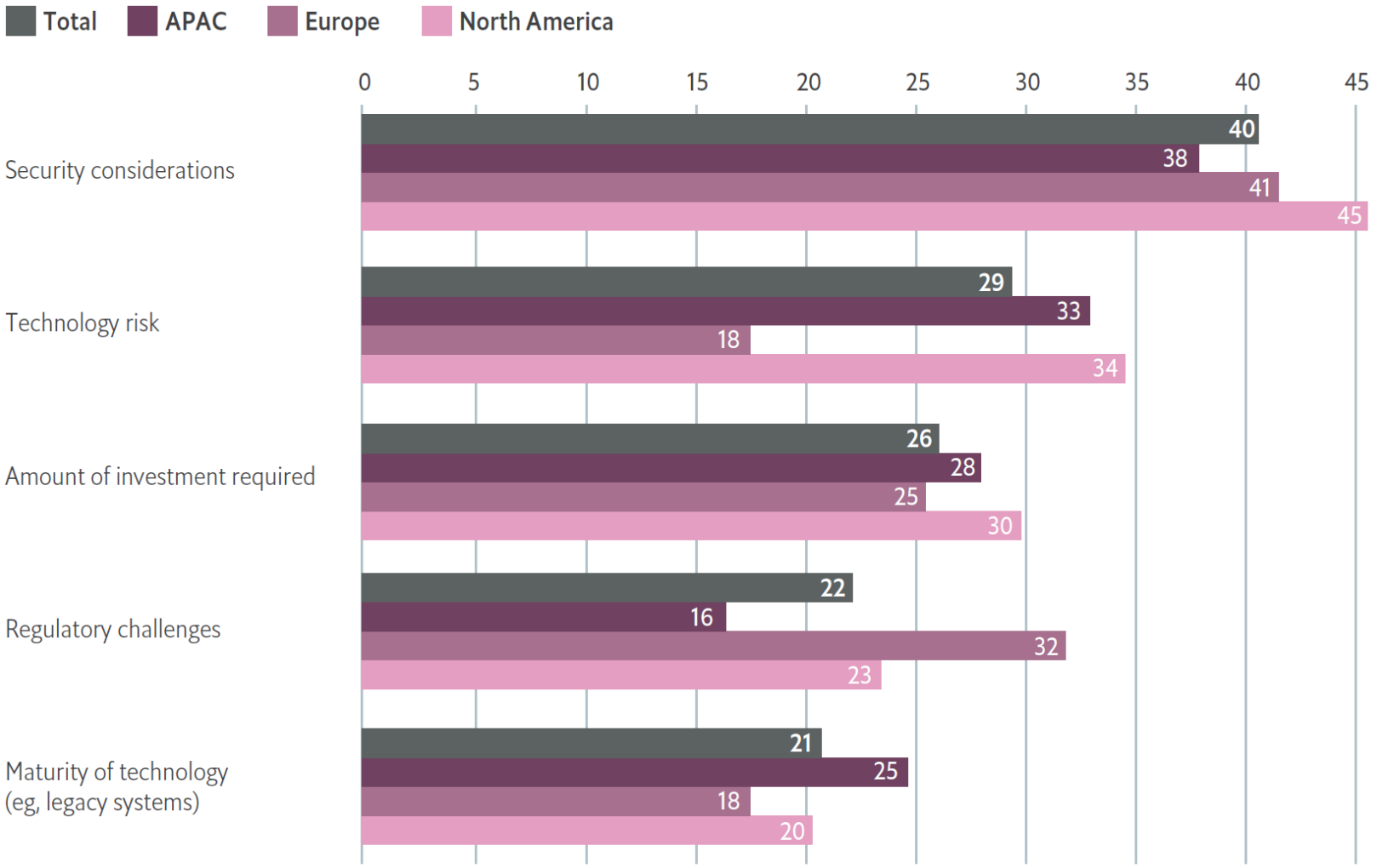
Collaboration with MITRE, Academic

Extensive literature survey

Taxonomy and vocabulary of adversarial machine learning, NISTIR 8269

Survey of Finance executives

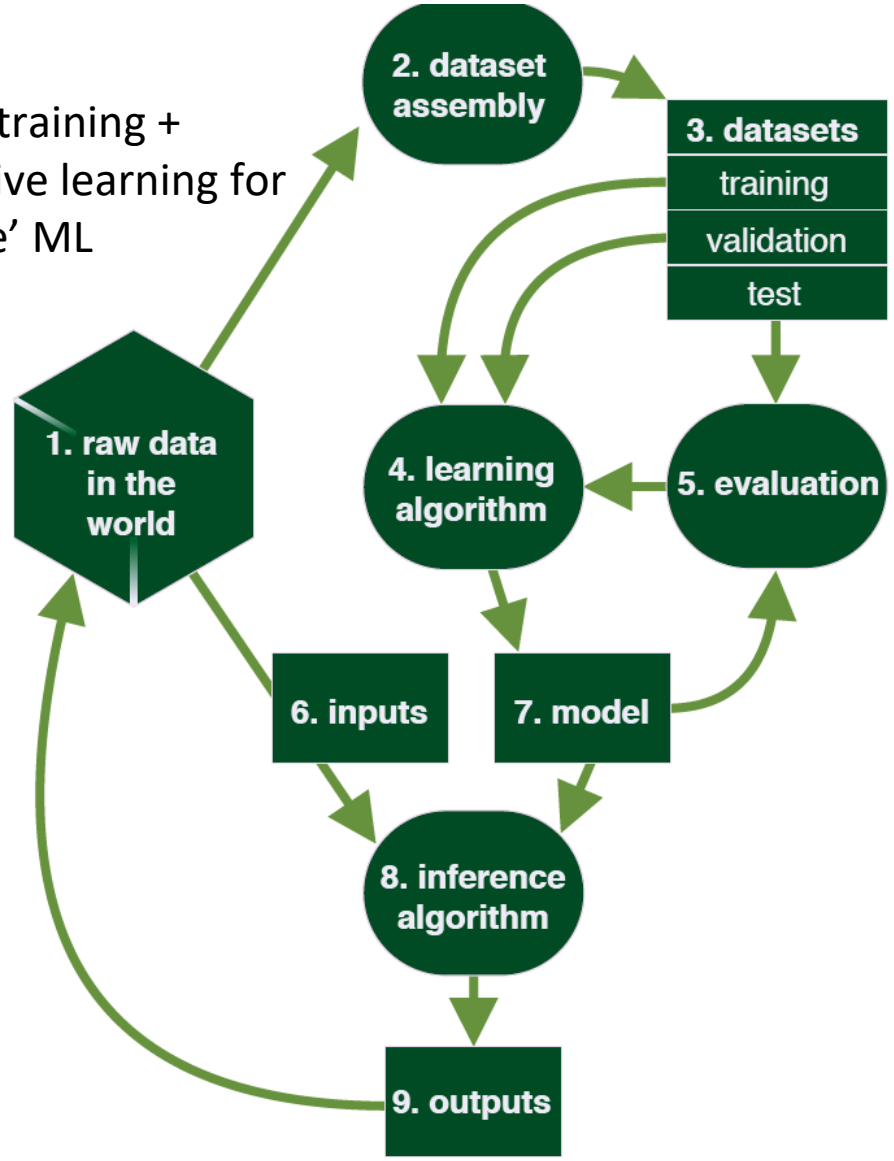
Figure 7: Principal risks
In your opinion, what are the principal industry risks of AI adoption?
(% of respondents)



Source: The Economist Intelligence Unit

The Economist Intelligence Unit: September 2020 report on AI in Financial Services

Initial training +
Adaptive learning for
'online' ML



Top 10 Risks

- Adversarial examples (1)
- Data poisoning (2)
- Online (in-use) system manipulation (1,2,3,4,5,7)
- Transfer learning attacks (5,4,7,8)
- Data confidentiality
- Data trustworthiness
- Reproducibility
- Overfitting
- Encoding integrity (5,4,7,8)
- Output integrity (8,9)

Credit:
An Architectural Risk Analysis of Machine Learning Systems: Towards More Secure Machine Learning
McGraw, Figueroa, Shepardson & Bonnet
https://berryvilleiml.com/docs/ara.pdf

Figure 1: Components of a generic ML system. Arrows represent information flow.

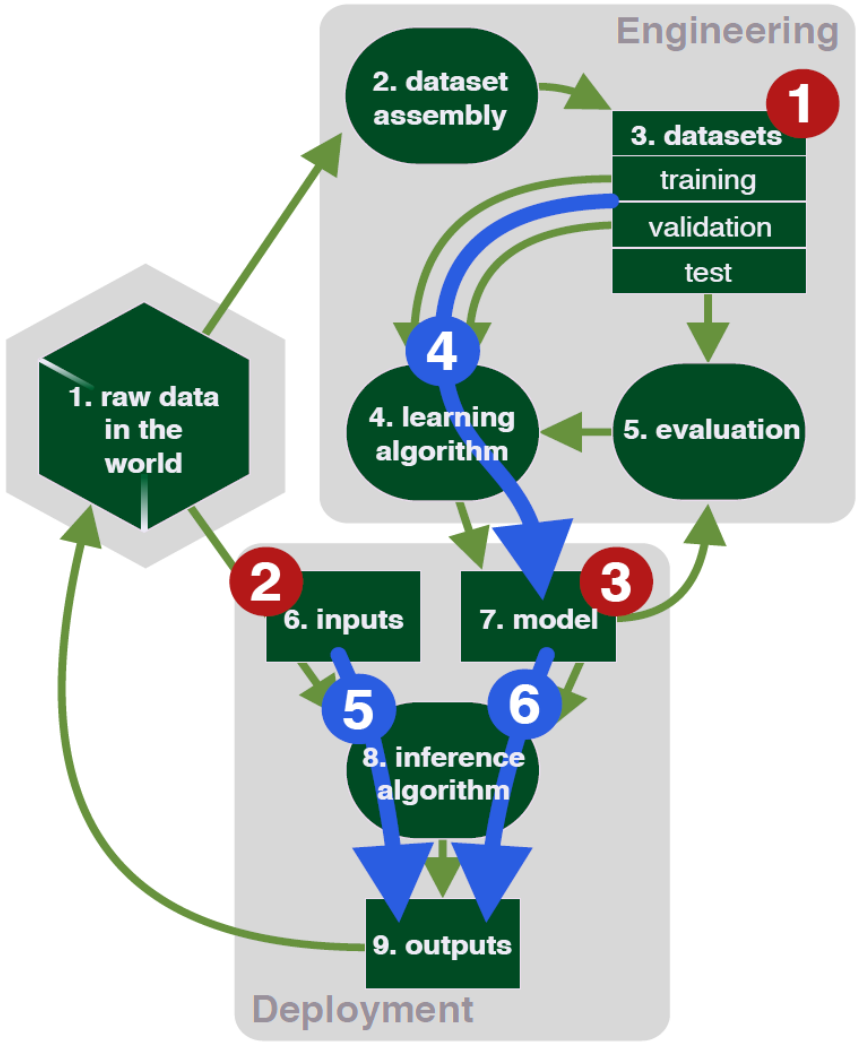


Figure 2: Known attacks and attack surfaces on ML systems. Manipulation attacks are pictured in red at the site of attack: (1) data manipulation. (2) input manipulation. (3) model manipulation. Extraction attacks are pictured in blue, showing the flow of information: (4) data extraction. (5) input extraction. (6) model extraction. Attack surfaces roughly correspond to gray plates:

Known attacks (from McGraw et al):

Manipulation attacks

1. Data manipulation
2. Input manipulation
3. Model manipulation

Extraction attacks

4. Data extraction
5. Input extraction
6. Model extraction

Explainable AI

Forms the bases of addressing fairness, bias, transparency, security, safety and ultimately trust in AI systems.

Developed principles of explainable AI

Socialize with experts in the community

Explainability, NISTIR 8312 draft

History

- Shannon, Turing
- Dartmouth 4, 1956
- Expert system, rules based, neural net, image net

Issues

- Weak vs **strong**
- Little notion of causality in statistics
- Uneven distribution of skills, most academics went to the usual suspects,
- Bias, explainability
- Skills shortage
- Deep Fakes
- Policy Makers, Regulators, Standards Bodies
- Hard problems are still hard
- Security
- Common sense still not common

Deep Fakes

History

How it works

What to do

Deep fake slides with permission from Dr Jan Kietzman. Note slideshare.net link in references

What are deepfakes?

"Deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content with a high potential to deceive" (Kietzmann et al. 2020)

Example: Rowan Atkinson (Mr. Bean) unexpectedly stars in a perfume commercial (original recorded with Charlize Theron).

View the original advert here:
https://youtu.be/VqSl5mSJXJs

View the deepfake here:
https://youtu.be/tDAToEnJEY8

What are deepfakes?

The phenomenon gained its name from a user of the platform

This person shared the first deepfakes by placing unknowing celebrities into adult video clips. This triggered widespread interest in the Reddit community and led to an explosion of fake content.

The first targets of deepfakes were famous people, including actors (e.g., Emma Watson and Scarlett Johansson), singers (e.g., Katy Perry) and politicians (e.g., Persident Obama).

Deepfakes matter because:

Believability: If we see and hear something with our own eyes and ears, we believe it to exist or to be true, even if it is unlikely.

The brain's visual system can be targeted for misperception, in the same way optical illusions and bistable figures trick our brains.

Accessibility: The technology of today and tomorrow will allow all of us to create fakes that appear real, without a significant investment in training, data collection, hardware and software.

Zao, the popular Chinese app for mobile devices lets users place their faves into scenes from movies and TV shows, for free.

How do deepfakes work?

Many deepfakes are created by a three-step procedure:

Step 1: The image region showing original person's face is extracted from an original frame of the video. This image is then used as input to a deep neural network (DNN), a technique from the domain of machine learning and artificial intelligence.

Step 2: The DNN automatically generates a matching image showing someone else instead of the original person.

Step 3: This generated face is inserted into the original reference image to create the deepfake.

How do deepfakes work?

The main technology for creating deepfakes is **deep learning**, a **machine learning** method used to train **deep neural networks (DNNs).**

DNNs consist of a large set of interconnected artificial neurons, referred to as units.

Much like neurons in the brain, while each unit itself performs a rather simple computation, all units together can perform complex nonlinear operations.

In case of **deepfakes**, this is mapping from one person to another.

How do deepfakes work?

Deepfakes are commonly created using a specific deep network architecture known as autoencoder.

Autoencoders are trained to recognize key characteristics of an input image to subsequently recreate it as their output. In this process, the network performs heavy data compression.

Autoencoders consist of three subparts:

- An **encoder** (recognizing key features of an input face).
- A **latent space** (representing the face as a compressed version).
- A **decoder** (reconstructing the input image with all detail).

Enterprise advice

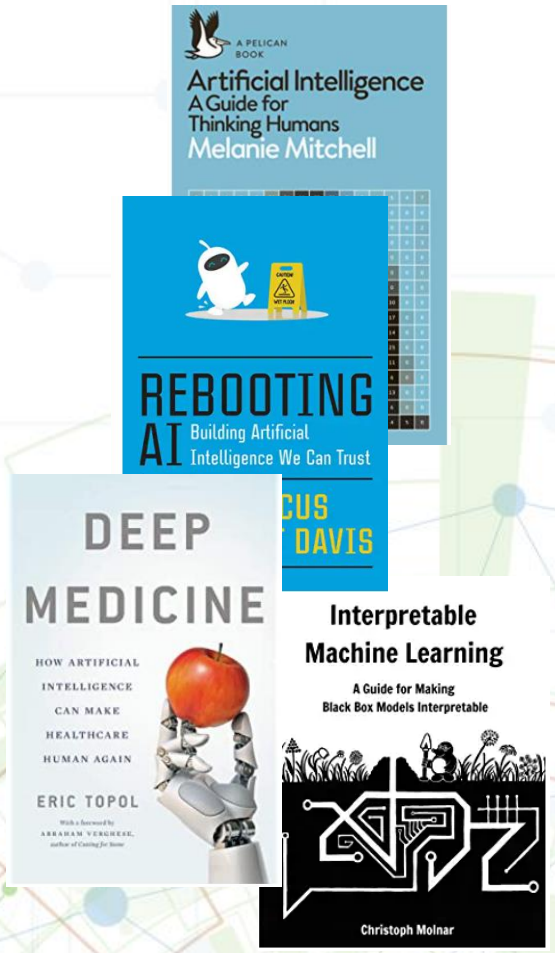
- Invest in learning and applying AI. Gain proficiency. Focus on business goals and problems. You need domain experience too.
- No ‘bet the farm’ projects - defined problems with business goals.
- Taking on very hard problems with technology you are not yet proficient in might work, probably not.

Policy, Regulation, Standards

- European Union, WEF
- OECD
- IEEEE. https://standards.ieee.org/initiatives/artificial-intelligence-systems/index.html
- ISO. https://www.iso.org/committee/6794475.html
- Sector vs global
- Context matters, autonomy, medical, military, financial, government, justice, transportation
- Thorny issues, facial recognition, autonomous vehicles, role of algorithms, bias, social justice
- At NIST, we respect role of regulators. Our lane is open broad, voluntary consensus bodies.

References (books)

- Artificial Intelligence: A Guide for Thinking Humans (2019), Pelican Books, Melanie Mitchell
- Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again (2019), Eric Topol
- Interpretable Machine Learning: A guide for making black box models interpretable (2019), Christopher Molnar
- Rebooting AI: Building Artificial Intelligence We Can Trust (2019), Gary Marcus and Ernest Davis



References – II (academia)

1. Towards the Science and Security of Privacy in Machine Learning
2. Security Evaluation of Pattern Classifiers Under Attack
3. Poisoning Attacks Against Support Vector Machines
4. Evasion Attacks against Machine Learning at Test Time
5. Towards Evaluating the Robustness of Neural Networks
6. Explaining and Harnessing Adversarial Examples
7. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing
8. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures
9. Ensemble Adversarial Training: Attacks and Defenses
10. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks
11. Defensive Distillation is Not Robust to Adversarial Examples
12. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks
13. Practical Black-Box Attacks against Machine Learning
14. Compression to the Rescue: Defending from Adversarial Attacks Across Modalities

References – III (media)

IBM, AI And The Battle For Cybersecurity, Forbes Sept 2020
https://www.forbes.com/sites/tiriasresearch/2020/09/17/ibm-ai-and-the-battle-for-cybersecurity/#3ef9e887a438

GPT-3: The First Artificial General Intelligence? Towardsdatascience Jul 2020
https://towardsdatascience.com/gpt-3-the-first-artificial-general-intelligence-b8d9b38557a1

3 ways criminals use artificial intelligence in cybersecurity attacks, TechRepublic Oct 2020
https://www.techrepublic.com/article/3-ways-criminals-use-artificial-intelligence-in-cybersecurity-attacks/

https://www.slideshare.net/IanMcCarthy/deepfakes-trick-or-treat

https://www.oecd.org/going-digital/ai/principles/

https://www.oecd.ai/dashboards?utm_campaign=The%20Batch&utm_medium=email&_hsmi=111648800&_hsenc=p2ANqtz--ZR-B0Ekm1ug4KTHIhF6EY6qylf1ZQfv3_LJPOb1_XPZM_OvCZVUGjAY6DzkAlpHXDrhq-9X_UPiBMbd-PKv0ewg_qdg&utm_content=111648800&utm_source=hs_email

References – V (NIST)

NIST Asks A.I. to Explain Itself
https://www.nist.gov/news-events/news/2020/08/nist-asks-ai-explain-itself
NISTIR 8312 Four Principles of Explainable Artificial Intelligence
https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf
Bias in AI Workshop (Aug 18 2020)
https://www.nist.gov/news-events/events/2020/08/bias-ai-workshop
https://www.nist.gov/topics/artificial-intelligence

Taxonomy and vocabulary of adversarial machine learning, NISTIR 8269
https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

https://www.nist.gov/itl/iad/mig/open-media-forensics-challenge
https://www.nist.gov/publications/nist-media-forensic-challenge-mfc-evaluation-2020-4th-year-darpa-medifor-pi-meeting

Discussion questions

- Well? - is it friend or foe?

Adversarial AI

Data poisoning

Bias

Reproducibility

Explain-ability/Transparency

Regulation

Ethics

AI Tech Power base

What are the opportunities?

What skills are needed?

What's the same?

- Risk view
- Patching?
- Red team?
- ??