



# **Trustworthy AI: Motivation and definitions**

Explainable and Trustworthy AI

Eliana Pastor

# Machine learning models are pervasive



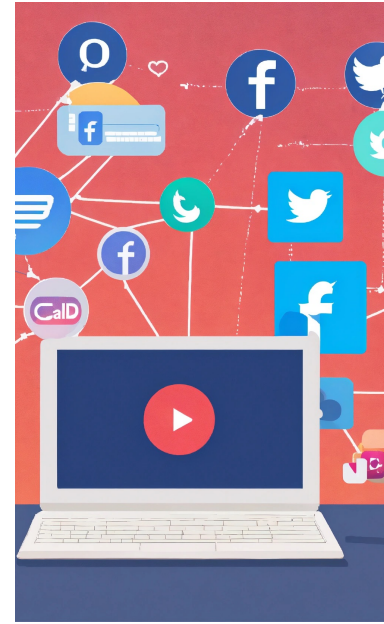
Finance



Medical  
diagnosis



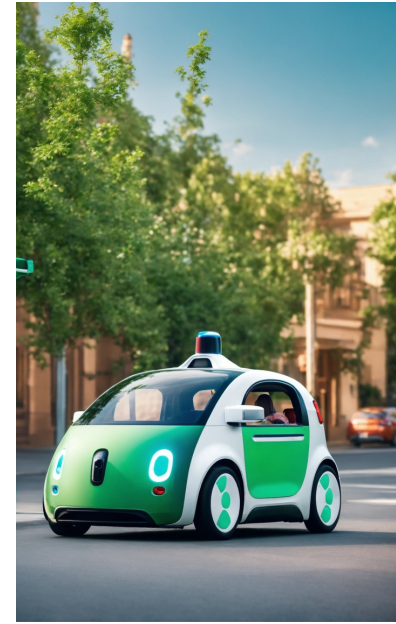
Recommender  
systems



Social  
network



Legal



Smart cities

and more..

# Can we trust these models?

A few examples..

- Models can learn true patterns.. but dangerous and potentially fatal if deployed
  - Pneumonia risk case
- Models can learn unfair and discriminatory patterns..
  - COMPAS – predict recidivism
  - Recruitment case
- Models can be fooled
- Models can make mistakes.. Are they accountable?
- Models can perpetuate historical biases + we don't trust them
  - Exam score prediction
  - Credit score prediction
- Not easy to make models fair

# Trust – Call for transparency

Increasing adoption of AI for medical diagnosis  
Accurate results, even ‘outperforming doctors’

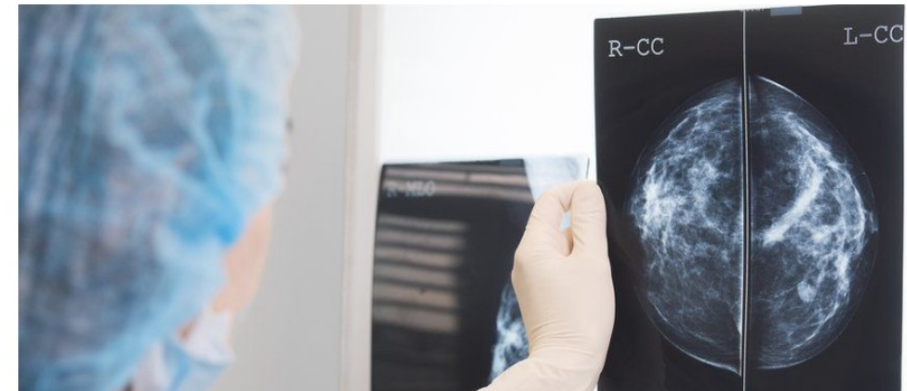
But...

## AI 'outperforms' doctors diagnosing breast cancer



**Fergus Walsh**  
Medical correspondent  
@BBCFergusWalsh

© 2 January 2020



Article | [Published: 01 January 2020](#)

## International evaluation of an AI system for breast cancer screening

[Scott Mayer McKinney](#) , [Marcin Sieniek](#), [Varun Godbole](#), [Jonathan Godwin](#)

[+ Show authors](#)

[Nature](#) **577**, 89–94 (2020) | [Cite this article](#)

**98k** Accesses | **1288** Citations | **3927** Altmetric | [Metrics](#)

# Trust – Call for transparency

Response article

Researchers call for **transparent and reproducible AI research**

Matters Arising | [Published: 14 October 2020](#)

## **Transparency and reproducibility in artificial intelligence**

[Benjamin Haibe-Kains](#) , [George Alexandru Adam](#), [Ahmed Hosny](#), [Farnoosh Khodakarami](#), [Massive Analysis Quality Control \(MAQC\) Society Board of Directors](#), [Levi Waldron](#), [Bo Wang](#), [Chris McIntosh](#), [Anna Goldenberg](#), [Anshul Kundaje](#), [Casey S. Greene](#), [Tamara Broderick](#), [Michael M. Hoffman](#), [Jeffrey T. Leek](#), [Keegan Korthauer](#), [Wolfgang Huber](#), [Alvis Brazma](#), [Joelle Pineau](#), [Robert Tibshirani](#), [Trevor Hastie](#), [John P. A. Ioannidis](#), [John Quackenbush](#) & [Hugo J. W. L. Aerts](#)

[Nature](#) **586**, E14–E16 (2020) | [Cite this article](#)

**15k** Accesses | **54** Citations | **520** Altmetric | [Metrics](#)

‘[...] The lack of details of the methods and algorithm code undermines its scientific value. Here, we identify obstacles that hinder **transparent and reproducible AI research** and provide solutions to these obstacles with implications for the broader field.’

# Trust – Predicting pneumonia risk case

Target: build a model to predict the risk of death in patients with pneumonia

Data from hospitalized patients

Created two models:

- a model they could **interpret**, less accurate
- a model they could not interpret, more accurate

They opted for the interpretable one, even if less accurate

**Need to understand** the reasons behind prediction



# Trust - Predicting pneumonia risk case

The interpretable model learned this association:

**history of asthma → lower chance of dying from pneumonia**

Counterintuitive!

Asthma is considered as serious risk factor for people who get pneumonia

Being interpretable revealed this is a **true pattern in the data**

- Asthmatics probably more attention, notice earlier the symptoms of pneumonia
- As high-risk patients, they get high-quality treatment sooner than other people
- Asthmatics actually have lower risk of death compared to the overall population
- But... using this model for deciding how admitting would be fatal and hurt asthmatics

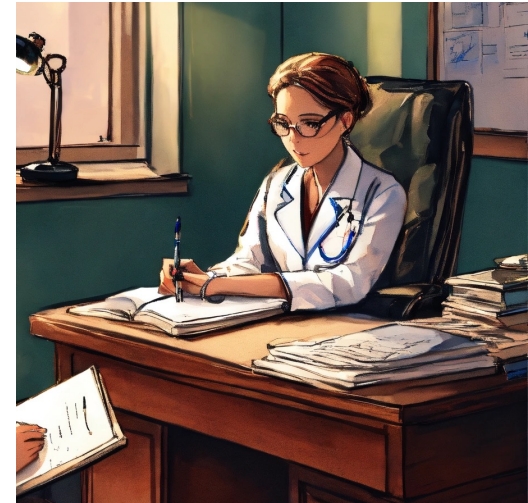
# Trust - Predicting pneumonia risk case

Experts could identify this issues since they could inspect the model

Using the non-interpretable model, this issue would have not been uncovered

The non-interpretable, while more accurate, could learn other dangerous patterns hidden in the data

In some high-risk applications as healthcare, it is **imperative for domain experts to analyze the model to understand** its behavior and decide if we can **trust** the model and use it





# Models can learn unfair and discriminatory patterns - COMPAS Case - Predicting recidivism

- Risk assessment tools can assist judges to make informed decisions, e.g., COMPAS score for risk of recidivism score
- Journalists of ProPublica analyze the data of 7,000 people arrested in Broward County, Florida, in 2013 and 2014
- Compared the risk scores assigned with the 'actual recidivism', i.e., individuals were charged with new crimes in a period of two years



## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

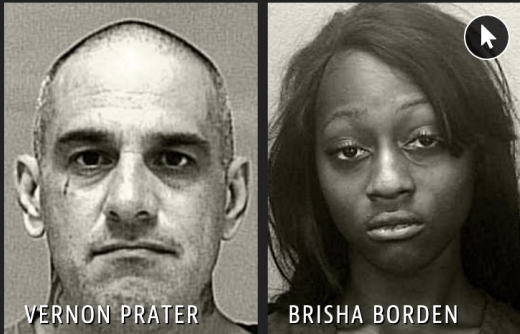
*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*

May 23, 2016

# COMPAS Score Case – Predicting recidivism

- Analysis revealed significant racial disparities
  - The algorithm wrongly assigned African-American defendants with a high risk of recidivism. The false positive are at almost twice the rate as white defendants
  - White defendants were mislabeled as low risk more often than black defendants

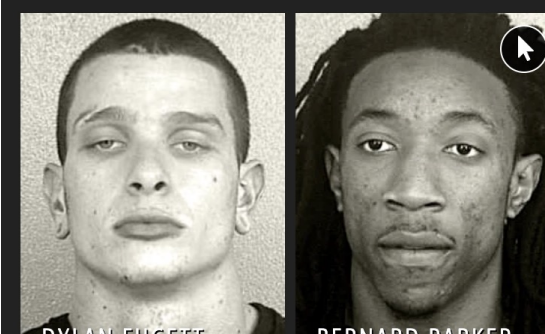
Two Petty Theft Arrests



|               |               |
|---------------|---------------|
| VERNON PRATER | BRISHA BORDEN |
| LOW RISK 3    | HIGH RISK 8   |

*Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.*

Two Drug Possession Arrests



|              |                |
|--------------|----------------|
| DYLAN FUGETT | BERNARD PARKER |
| LOW RISK 3   | HIGH RISK 10   |

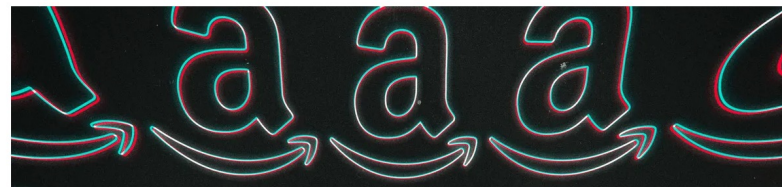
*Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.*

# Models can learn unfair and discriminatory patterns – Recruiting tool

- AI recruiting tool showed bias against women
  - Penalized applicants who attended all-women’s colleges and resumes that contained the word “women’s” (e.g., “women’s chess club”).
- AI systems learn to make decisions by looking at historical data. Hence, they can perpetuate existing biases
  - Bias: tech is a male-dominated working environment

TECH / AMAZON / ARTIFICIAL INTELLIGENCE

## Amazon reportedly scraps internal AI recruiting tool that was biased against women

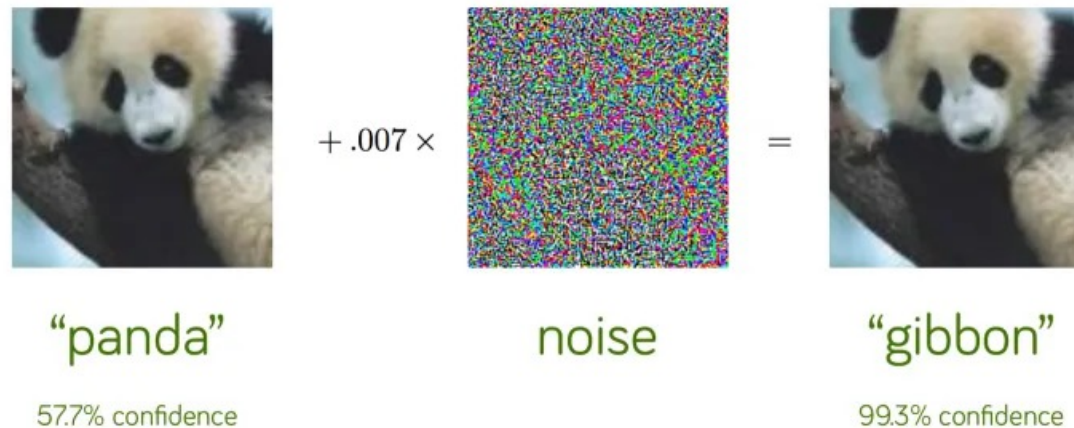


/ The secret program penalized applications that contained the word “women’s”

# Models can be fooled - I

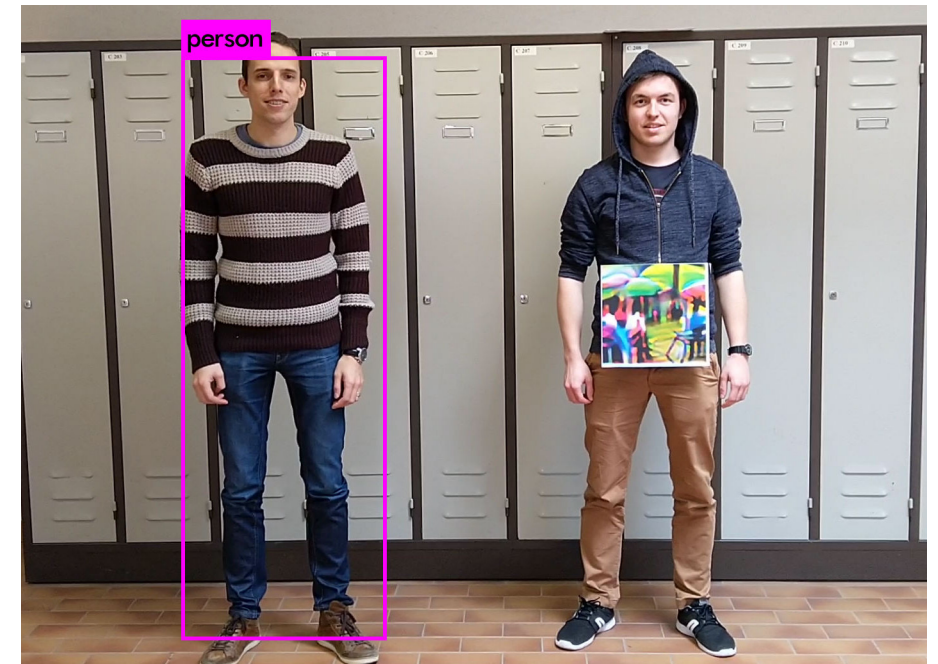
In 2015, researchers found that it was easy to fool CNNs by carefully varying input. By adding some noise, the classifier identifies an image of class to a completely different one

- In the example, it classifies an image of a panda as a gibbon with an accuracy of over 99%.
- For us as humans, both the images are easily identifiable as pandas.



# Models can be fooled -II

- Researchers created an adversarial patch that can hide persons from a person detector (YOLOv2)
  - The person without a patch is successfully detected
  - The person holding the patch is ignored
- Risks & Attacks
  - It can be used maliciously to circumvent surveillance systems



Thys, Simen, Wiebe Van Ranst, and Toon Goedemé. "Fooling automated surveillance cameras: adversarial patches to attack person detection." *IEEE/CVF workshops*. 2019.

# Models can make mistakes.. Are they accountable?

- A Canadian customer reportedly asked the chatbot for clarification on refunding
  - Chatbot provided fake information on the refund policy

- The airline argued "the chatbot is a separate legal entity that is responsible for its own actions"
- The court sided with the customer, forcing the company to issue the passenger a partial refund and pay for his court fees.
  - The airline shut down the chatbot

## **Air Canada Must Honor a Fake Refund Policy Created by Its Chatbot, Court Says**

The airline argued that the chatbot should be responsible for its own actions.



By [Emily Price](#)

Updated February 18, 2024



<https://www.pcmag.com/news/air-canada-must-honor-a-fake-refund-policy-created-by-its-chatbot-court>  
Court order: <https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html>

# Models can perpetuate historical biases + we don't trust them

## UK ditches exam results generated by biased algorithm after student protests



By [Jon Porter](#), a reporter with five years of experience covering consumer tech releases, EU tech policy, online platforms, and mechanical keyboards.

Aug 17, 2020, 6:16 PM GMT+2

- Context: A-levels are final exams taken before university. They have a huge impact on which institution students attend
  - During 2020 pandemic, exams were not taken.
- Teachers asked to predict the grades – then adjusted via an algorithm
- Accusations that the system was biased against students from poorer backgrounds
- The major issue was lack of transparency of the system, how predictions were made
  - No trust in the system

# Models can be biased .. + we don't trust them

- Husband and wife applied to Apple Card
  - Same credit history but.. the husband got 20 times credit limits than his wife
- Impact on brand reputation
- Investigations if gender discrimination

## Apple co-founder says Apple Card algorithm gave wife lower credit limit

By Subrat Patnaik

November 11, 2019 2:10 AM GMT+1 · Updated 4 years ago

Aa

## *Apple Card Investigated After Gender Discrimination Complaints*

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.



# Models can be biased .. + we don't trust them

- After investigations.. They found there was no bias
- The inputs were not actually the same.

**The Apple Card doesn't actually discriminate against women, investigators say**

By [Ian Carlos Campbell](#)

Mar 24, 2021, 1:34 AM GMT+1

- The investigators indicated that model did not consider protected/prohibited characteristics of applicants and would not produce disparate impacts
- Still, trust is difficult to rebuilt

# Not easy to make models fair

Gemini image generation got it wrong.  
We'll do better.

Feb 23, 2024  
2 min read

We recently made the decision to pause Gemini's image generation of people while we work on improving the accuracy of its responses. Here is more about how this happened and what we're doing to fix it.

- Google blocked the ability to generate images of people on Gemini after some users accused it of anti-White bias
  - 'We did not want Gemini to refuse to create images of any particular group. And we did not want it to create inaccurate historical — or any other — images.'



End Wokeness ✓

@EndWokeness

America's Founding Fathers, Vikings, and the Pope according to Google AI:

[Traduci post](#)

Certainly! Here is a portrait of a Founding Father of America:



Sure, here is an image of a Viking:



Sure, here is an image of a pope:



# On the need of Trustworthy AI

- Society increasingly relies on AI for critical decisions in healthcare, finance, justice, and more
- All these examples highlight the need to build AI models that we can trust
- The need is highlighted by regulations and guidelines
  - EU AI Act: first regulation on artificial intelligence
  - GDPR
  - EU's Ethics guidelines for trustworthy AI

# Key requirements for Trustworthy AI

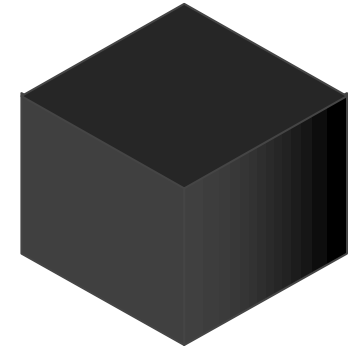
1. Transparency & Explainability
2. Technical robustness and safety
3. Fairness, diversity and non-discrimination
4. Accountability
5. Privacy and data governance
6. Human Agency and Oversight
7. Societal and environmental well-being





# 1. Transparency & Explainability

Most AI models are black boxes



Opaqueness of a model can be at multiple levels, involving

- Data
- Model/System/Algorithm
- Learned function/pattern → reasons behind its functioning
- Intention and business model of the AI product

These elements should be **transparent** – clear, disclosed - to the end users

This is achieved through: **Explainability, Traceability and Communication**



# 1. Transparency & Explainability

## Explainability

- Ability to explain the **reasoning** behind the decisions or predictions that the **AI system makes** in understandable terms **to humans**
- Ability to explain the technical processes of the AI system
- **Tailored explanations.**
  - AI systems and their decisions should be **explained** in a manner **adapted** to the involved stakeholder (e.g. layperson, domain experts, regulator or AI researcher).



# 1. Transparency & Explainability

## Explainability

- **Meaningful information and the *'right for explanation'*.**
  - AI driven decisions must be explained to and understood by those directly and indirectly affected
  - If AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process.

Articles 13 and 14 of the GDPR state that, when profiling takes place, a data subject has the right to *"meaningful information about the logic involved."*



# 1. Transparency & Explainability

## **Explainability**

- **Trade-off accuracy-explainability**
  - Consider the trade-offs between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)
- **Ensuring business model transparency**
  - Provide explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it





# 1. Transparency & Explainability

## **Traceability**

- The data sets and the processes that yield the AI system's decision should be documented to increase transparency
  - Including the data collection, data labelling, algorithm used
- Traceability facilitates auditability and explainability.



# 1. Transparency & Explainability

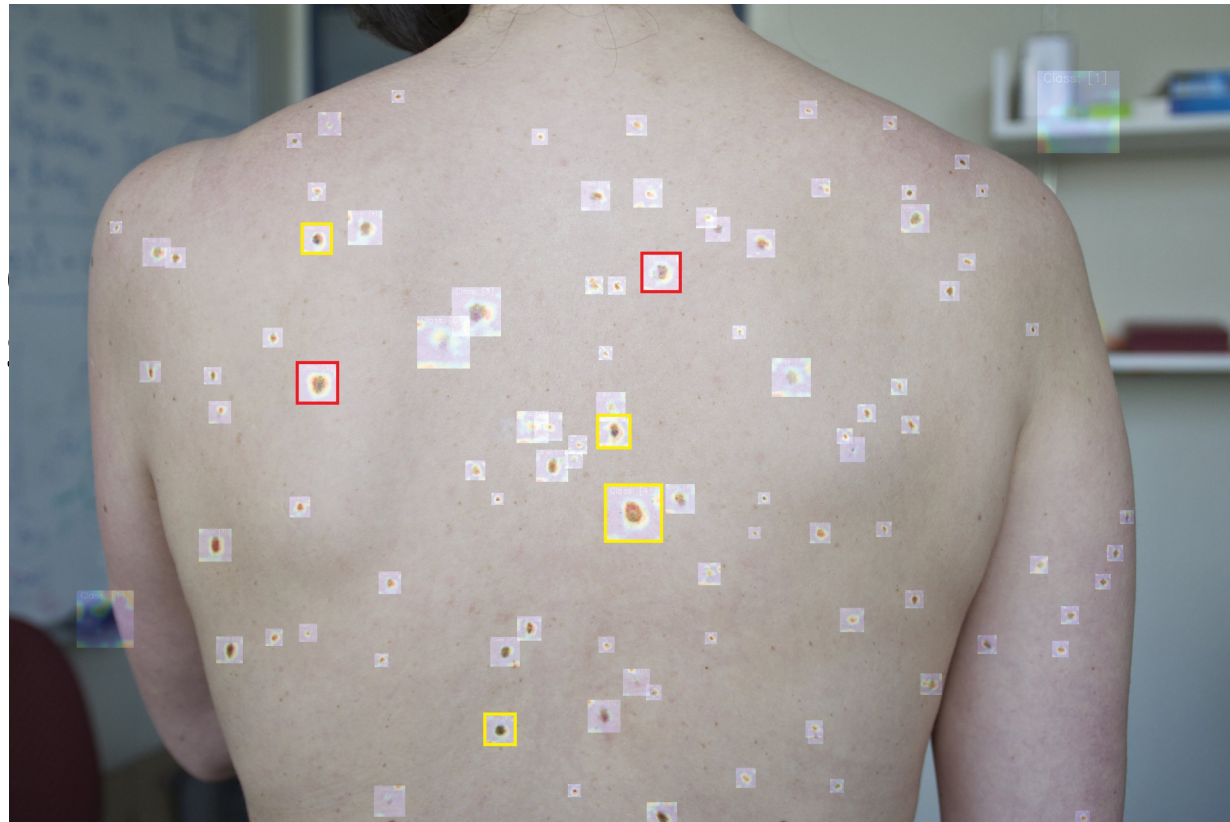
## Communication

- Communicate the AI system's capabilities, benefits, limitations and potential risks to practitioners or end-users
- Humans need to be aware that they are interacting with an AI system
- Humans have the right to be informed that they are interacting with an AI system
- Provide appropriate training material and educate the users on how to adequately use the AI system



# Example - Transparency & Explainability

Medical diagnosis AI system designed to assist doctors in diagnosing skin cancer based on images





# Examples for 1. Transparency & Explainability

- AI system should provide explanations for its diagnostic decisions in understandable terms
  - to healthcare professionals - offering more technical details for healthcare professionals
  - and patients – simplified explanation, clear
- Assess trade-off: e.g., build a complex black model and explaining with it with post-modelling explainability techniques
- Outlines the capabilities, benefits and limitations. Inform the patient a system is used

Image (melanoma)

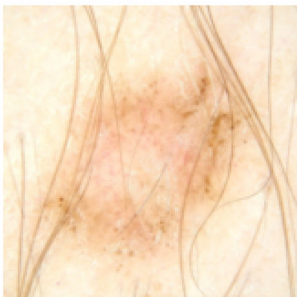


Image (benign)

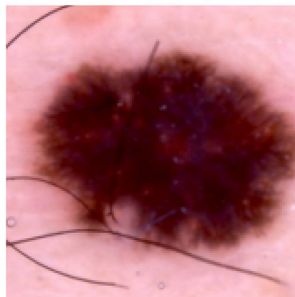
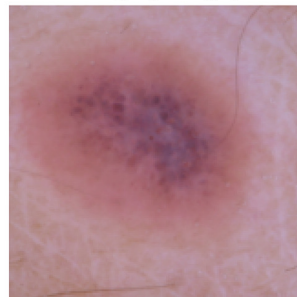
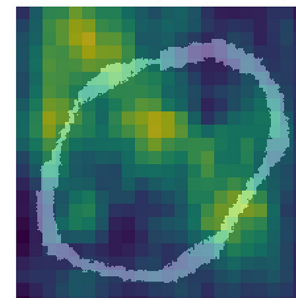


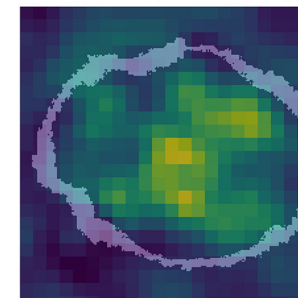
Image (benign)



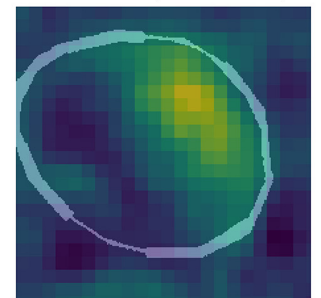
Occlusion (iAUC=.442)



Occlusion (iAUC=.525)



Occlusion (iAUC=.813)





## 2. Technical robustness and safety

### AI systems

- need to be resilient and secure
- should be developed with a preventative approach to risks
- behave as intended while minimizing unintentional and unexpected harm and preventing unacceptable harm

This is achieved by addressing:

General Safety, Resilience to Attack and Security, Accuracy, Reproducibility



## 2. Technical robustness and safety

### General Safety

- Define potential risks associated with the use of AI system across various application area
  - Included definition of metrics on how to assess risks
  - Define a process for measuring and assess risks
  - Inform end-users of potential risks
- Identify possible threats (design faults, technical faults)
  - Included risks of malicious use, misuse or inappropriate use
- Define a fallback plan in case of problems
  - Included statistical or rule-based procedure or feedback from a human operator



## 2. Technical robustness and safety

### **Resilience to attack and security**

- Protect AI systems against vulnerabilities
- Identify possible attacks to prevent them
  - Attacks may come from various levels: on the data (data poisoning, manipulation of the training data), on the model (model leakage, model inversion to infer the parameter), adversarial attacks to change or control the system behavior (model evasion)
- Put in place measures to ensure the integrity, robustness and overall security
  - Monitor the system



## 2. Technical robustness and safety

### **Accuracy**

AI systems should be accurate, able to make correct predictions, recommendations, or decisions based on data or models

- Perform a well-formed development and evaluation process to support, mitigate and correct unintended risks from inaccurate predictions
  - Ensure data are up-to-date, high quality, complete and representative
- Consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects
- Monitor AI performance and document it
  - AI systems can make mistakes... the system should indicate how likely these errors are
- Requirement of high level of accuracy for critical application, affecting directly human lives





## 2. Technical robustness and safety

### **Reliability and Reproducibility**

AI systems and their results should be

- Reliable: work properly with a range of inputs and in a range of situations
- Reproducible: same behaviour when repeated under the same conditions
- Document and operationalize processing for testing and verifying reliability and reproducibility
  - Describe the data, systems and pipeline for its development and deployment
- Monitor the system to ensure its reliability
  - Control the correctness of AI system operation under the conditions of expected use and over time



# Example for Technical robustness and safety

- Potential Risks: Misdiagnosis leading to delayed treatment
- Technical Faults: problems in the image processing may cause the system to misinterpret features of skin lesions
- Inappropriate Use: non-experts attempt to interpret diagnostic results without the necessary training
- Security: prevent unauthorized access, encryption of data
- Accuracy: extensive validation using diverse datasets, including cases with varying skin types, lesion sizes, and stages of cancer
- Error Indication: indicate likelihood of errors





# 3. Fairness, diversity and non-discrimination

Data reflects biases and discriminations of our society

- As a consequence, AI systems may encode biases in the data..
  - inadvertent inclusion of historic bias, incomplete and non-representative data
- ML can therefore perpetuate such biases
  - unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalization



# 3. Fairness, diversity and non-discrimination

## Avoid Unfair Bias

- Identify possible discriminatory bias and remove them
  - At multiple levels, such as data collection, data processing, algorithm design
- Assess and enforce diversity and representativeness in the data
  - Evaluate systems for multiple groups or use cases
- Evaluate the fairness of the systems
  - Clearly define the fairness evaluation measures
  - Define mechanisms to ensure fairness of the systems
- Encourage including experts from diverse backgrounds and disciplines to ensure diversity of opinions



# 3. Fairness, diversity and non-discrimination

## **Accessibility and Universal Design**

AI systems should be designed for users regardless of their age, gender, abilities or characteristics

Universal Design principles to addressing the widest possible range of users



# Example - Fairness, diversity and non-discrimination

- Enforce Representativeness in the Data: collect data from lighter and darker skin tones
- Evaluate Fairness: assess the performance across different demographic groups, including age, gender, and ethnicity. Identify disparities and actively work to address them
- Encourage Diversity of Expertise: dermatologists and data scientists from diverse backgrounds





# 4. Accountability

One is responsible for their action – and their consequences – and must be able to explain their aims, motivations, and reasons

**Who Holds the Accountability? Various entities, e.g.,**

- **AI Users, the individuals using AI systems:** Understand their functionality and potential limitation, ensuring appropriate use
- **Businesses employing AI:** Establish clear guidelines for its use. They are accountable for the consequences of AI use within their organisation
- **AI Developers:** They should ensure that the AI is designed and trained responsibly and with safety measures to prevent misuse or errors
- **Data Providers:** Data providers are accountable for the quality and accuracy of the data



# 4. Accountability

## **Auditability**

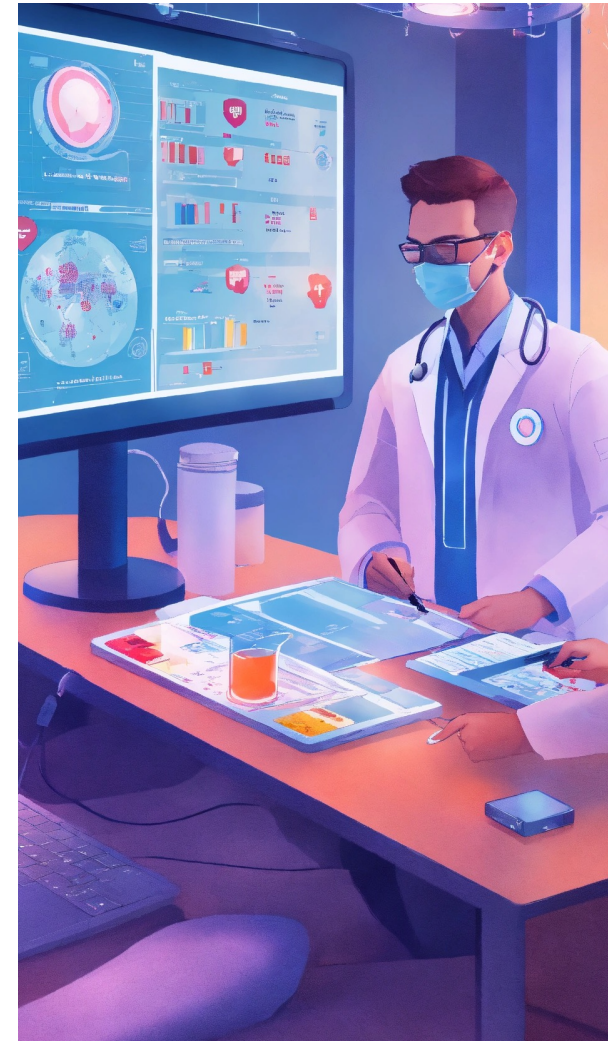
- Assessment of algorithms, data and design processes
  - Internal and external auditors
- Develop mechanisms to facilitate the auditability
  - e.g, via traceability, logging





# Example - Accountability

- Healthcare practitioners using the AI system are responsible for understanding its functionality and limitations.
- Data providers, such as medical institutions, are accountable for the quality and accuracy of the data used to train the AI system. Collaborate with dermatologists and domain experts to verify the integrity of the data
- Internal audit teams should periodically review the AI system's performance and adherence to established guidelines and protocols. External auditors may be engaged to provide independent assessments





# 5. Privacy and data governance

## Privacy

Fundamental right

- AI systems could infer private information (preferences, sexual orientation, age, gender, religious or political views)
- Assess the impact of the AI system on privacy and data protection
- Ensure privacy for the entire cycle
  - Information initially provided by the user and generated ones during its interaction
- Ensure data collected will not be used to unlawfully or unfairly discriminate
- Right to be forgotten



# 5. Privacy and data governance

## Data governance

Process of managing data during its entire life cycle

- Quality and integrity of data
  - Ensure data is secure, private, accurate, available, and usable
  - Ensure the quality, also respect to its representativeness and fairness
  - Prevent including malicious data
  - Test and document data used at each step such as planning, training, testing and deployment
- Access to data - outline how can access data and under which circumstances
- Ensure following data protection regulation (e.g., GDPR)



# Example - Privacy and data governance

The system collects various types of personal information from users, including images of skin lesions, demographic data (such as age and gender), and medical history.

- Users must be informed about the types of data collected, how it will be used, and their rights regarding its privacy and protection.
- User consent should be obtained before collecting any personal information
- User data should securely be encrypted during transmission and storage
- Access to user data is strictly controlled and granted only to authorized personnel





# 6. Human Agency and Oversight

## Human Agency

AI systems should **support** human decision-making

- The principle of user autonomy should be central to the system functionality
- Right not to be subject to a decision based solely on automated processing when AI systems produce legal effects on users or significantly affects them
  - Art. 22 GDPR: “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her”



# 6. Human Agency and Oversight

## Oversight

Ensure AI systems do not undermine human autonomy or causes other adverse effects

Governance mechanisms such as a

- human-in-the-loop (HITL): human intervention in every decision cycle
- human-on-the-loop (HOTL): human intervention during the design cycle of the system and monitoring the system's operation
- human-in-command (HIC): oversee the overall activity of the AI system and decide when and how to use the system



# Example - Human Agency and Oversight

- Goal is to support Human Decision-Making. The final diagnosis and treatment decisions are made by healthcare professionals, who use the predictions as supplementary info
- Users have the authority and autonomy to accept, modify, or reject the recommendations





# 7. Societal and environmental well-being

AI systems should benefit all human beings, including future generations.

- Ensure that AI systems are sustainable and environmentally friendly
  - Assess the (potential) positive and negative impacts on the environment
  - Examine of the resource usage and energy consumption during training
- Carefully consider the AI system's social and societal impact
  - E.g., Impact human work; effect on institutions and democracy
- Envisage actions to minimize potential societal harm of the AI system





# Example - Societal and environmental well-being

- The system should be designed to benefit all individuals by providing accessible and accurate diagnostic assistance.
  - We should ensure equal access for people from diverse socio-economic backgrounds and geographic locations
- Minimize Societal Harm. Identify and address them, ensuring that the system's recommendations do not inadvertently perpetuate biases, contributing to healthcare disparities
- The implementation of system is designed to complement the work of healthcare professionals rather than replace them



# References

- Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." *KDD* 2015.
- <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015
- Thys, Simen, Wiebe Van Ranst, and Toon Goedemé. "Fooling automated surveillance cameras: adversarial patches to attack person detection." IEEE/CVF workshops. 2019.
- Liang, W., Tadesse, G.A., Ho, D. *et al.* Advances, challenges and opportunities in creating data for trustworthy AI. *Nat Mach Intell* **4**, 669–677 (2022).
- Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* 55, 9, Article 177 (September 2023), 46 pages. <https://doi.org/10.1145/3555803>
- Bokadia, Harshit, et al. "Evaluating perceptual and semantic interpretability of saliency methods: A case study of melanoma." *Applied AI Letters* 3.3 (2022): e77.
- [www.med-technews.com/news/Digital-in-Healthcare-News/usw-researchers-work-on-app-to-streamline-skin-cancer-diagno/](http://www.med-technews.com/news/Digital-in-Healthcare-News/usw-researchers-work-on-app-to-streamline-skin-cancer-diagno/)