DS 517– Lecture 3

Transparency in AI & XAI

DS-517-50: Ethics and Bias in AI
2024 Fall
MONMOUTH CAMPUS
M 7:30 PM - 10:20 PM
9/3/2024 - 12/9/2024
Howard Hall, 309 LECTURE

Arup Das

adas@Monmouth.edu

Disclaimer:

- The views expressed are solely those of the presenter and not affiliated with any other party.
- This presentation is free of copyright violations, and external sources have been appropriately credited.
- T**he content within this presentation is legally protected; unauthorized reproduction, including photography, will result in legal action.**
- **This material is not intended for distribution and must remain solely within the confines of this class. Do not distribute slide or assignments to other students**
- Using cameras to take screenshots or photographs of the slides is strictly prohibited.

Date	Week	Class Format/Location/Time	Topics	Readings Required (Due before class)	Assignment/Quiz
September 9, 2024	Week_1	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Ethics & Human-Centered Design		
September 16,2024	Week_2	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Algorithms and Accountability	**Book 1 – Chapter 1, 3, 9**	**Assignment 1- Presentation on the professor assigned reading Due Sep 23,2024**
September 23,2024	Week_3	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AIA/Transparency in AI	Book 1 – Chapter 4	
September 30, 2024	Week_4	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Privacy, Security, and Inclusion	Book 1- Chapter 6, 7,8	**Assignment 2- Presentation on the professor-assigned reading Due Oct 7, 2024**
October 7, 2024	Week_5	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Fairness & Bias	Book 1- Chapter 2 Book 2 – Chapter 8	
October 14, 2024	Week_6	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Fairness & Bias	Book 2 – Chapter 8	**Assignment 3- Presentation on the professor-assigned reading Due Oct 21, 2024**
October 21, 2024	Week_7	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Regulatory Frameworks (US and Europe)	Professor Handout	
October 28,2024	Week_8	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Regulatory Frameworks (Continued)	Professor Handout	**Assignment 4 - Presentation on the professor-assigned reading Due Nov4,2024**
November 4, 2024	Week_9	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Explainable AI - Introduction to Model Interpretability	Book 2 – Chapter 1 ,2	
November 11, 2024	Week_10	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Explainable AI - Advanced Topics	Book 2 – Chapter 3, 4 ,5	**Coding Assignment – Due Nov 22, 2024**
November 18,2024	Week_11	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	Case Studies in Explainable AI	Book2 – Chapter 6,7	
November 25, 2024	Week_12	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20PM	AI Ethics in Practice	Book 2 – Chapter 9	
December 2, 2024 (Last class)	Week_13 &14	On-Premise/Howard Hall, 309 LECTURE/7:30 PM-10:20 PM	AI Ethics Case Essay Preparation/Final Exam and Course Wrap-Up		**Final Essay, Final Exam Due Dec 8, 2024, before midnight EST**

Course Logistics

1. OneDrive link for professor notes and assignments/quiz -
2. Check your Monmouth email for announcements
3. Check your Monmouth calendar for Zoom links for office hours and remote lectures
4. My contact information: adas@monmouth.edu, Cell # 917-523-7683
4. Office hours (zoom only) – Friday (EST)
5. Assignment submission to professoraruprdas@gmail.com (Notation for files: Assignment_1_Name_of_Student), Colab notebooks ipynb file and html file, all presentation in ppt format.
6. Quiz submission to professoraruprdas@gmail.com (Notation for file : Quiz_1_Name_of_Student. doc , Quiz_2_Name_of_Student.doc)

Lecture 2 Recap

What is an AIA?

What is an AIA? An Algorithmic Impact Assessment is a tool for identifying the potential societal impacts of an algorithmic system *before* it's launched.

AIAs are very much in the early stages of development, and as such there's no standard methodology on how to put one together yet. However, there's huge interest in tools like AIAs — they have so far been mostly proposed for public sector use, and there is already one 'live' example of an AIA tool being used in the Canadian government. So in these early stages, it's important to consider the potential use-cases for AIAs.

https://www.hattusia.com/post/accountability-in-ai-algorithmic-impact-assessments-jenny-brennan-lara-groves

ALGORITHMIC IMPACT ASSESSMENTS- NEPA

The National Environmental Policy Act is a United States environmental law that promotes the enhancement of the environment and established the President's Council on Environmental Quality

The NEPA model implies a highly detailed impact assessment, potentially running to hundreds of pages. It demands thorough answers to **open-ended questions that explain the design process.** Other features of the NEPA model are transparency and public participation via a notice and comment framework. Because transparency, and specifically notice and comment frameworks, are part of the regulation that is usually applied to the public sector in the United States, it is perhaps not surprising that these proposals tend to focus on the public sector, rather than the private sector.

https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf

ALGORITHMIC IMPACT ASSESSMENTS- GDPR/DPIA

AIA draws on European data protection law. Article 35 of the GDPR requires companies to perform DPIAs whenever data processing "is likely to result in a high risk to the rights and freedoms of natural persons.

- The DPIA envisions a similarly expansive scope of work to the NEPA model, including a "systematic description" of the processing, justifications, and plans for mitigation.
- One difference from the NEPA approach is that there is no explicit requirement to describe all the reasonable and rejected choices. The only requirement is to sys-tematically evaluate the actual program that is to go forward.
- In practice, however, the requirement to show all the "measures envisaged" to mitigate dangers might be broad enough to encompass the same idea.[122] The most significant difference is in transparency.
- Although the official guidance on DPIAs recommends making a summary of the DPIA public, publication — of even a summary — is not re-quired.[123] Instead DPIAs are performed in collaboration with member

https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf

ALGORITHMIC IMPACT ASSESSMENTS- Questionnaire Based

- The third approach is the one taken by the government of Canada. Under Canada's Directive on Automated Decision-Making, government agencies that use algorithmic decision making must complete an AIA both before production and before the project goes live. The AIA consists of "around 60 questions related to [] business process, data and system designed decisions."126

The questions touch on most of the topics people care about with respect to algorithms. Some of the questions go to the thoughts behind the process (e.g., "**What is motivating your team to introduce automation into this decision-making process**? (Check all that apply),"with choices related to backlog, efficiency, quality, and being innovative).Other questions ask about the stakes of the decisions, **the sec-tor, the degree of explanation or human involvement, and so on**. Each of these questions receive a point total. **That point total then determines whether the overall risk falls within one of four wide bands (Impact Levels I–IV),** and agencies implementing algorithmic system that fall within a given band must take certain increasingly involved remedial actions to mitigate the anticipated harms. While most of the questions are multiple choice, some do include written answers.[131] The written answers are not scored, but can be made public.[13]

https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf

Importance of AIA

- Early Intervention – Early stage interventions to inform projects before they are built. *(The NEPA and DPIA models require completion of the AIA before deployment of the project. The Canadian AIA is meant to be filled out before design and again after implementation)*

- Open Ended Questions - An effective AIA must ask open-ended questions, inviting bot-tom-up explanations. The algorithmic systems of interest are highly complex and far from fully understood

- Accountability – Legal requirement

https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf

Canada - Algorithmic Impact Assessment tool

Honour the memory of Her Majesty Queen Elizabeth II

Français

Government of Canada | Gouvernement du Canada

Search Canada.ca

MENU

Canada.ca > About government > Government in a digital age > Digital government innovation > Responsible use of artificial intelligence (AI)

Algorithmic Impact Assessment tool

On this page

Launch the AIA tool

permission of Arup Das

https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Topics

1. Algorithms and Accountability
2. Accountability Legislature
3. AIA – Algorithm Impact Assessment

Module 3 – What is Transparency Deep Dive

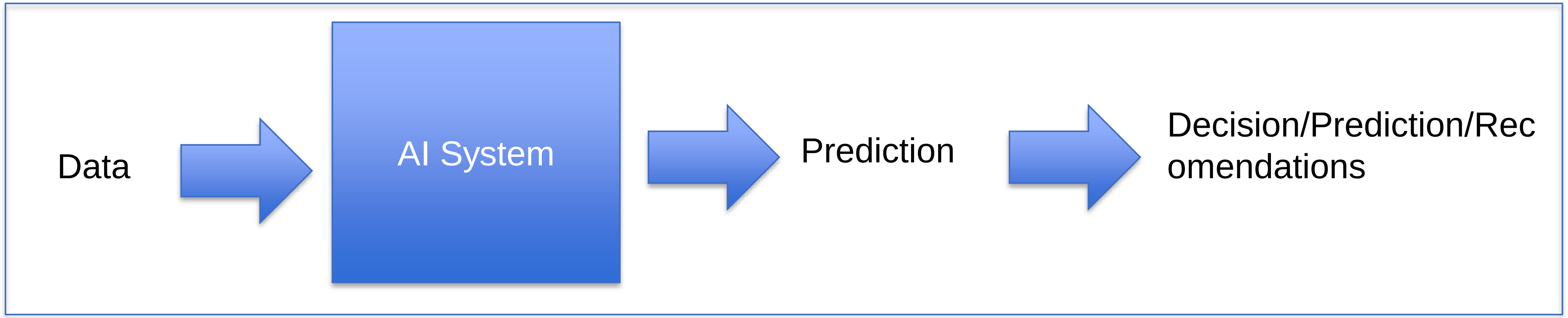
AI Transparency

The point of transparent AI is that the outcome of an AI model can be properly explained and communicated. "Transparent AI is explainable AI. It allows humans to see whether the models have been thoroughly tested and make sense, and that they can understand why particular decisions are made."

https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nl-innovation-bringing-transparency-and-ethics-into-ai.pdf

Transparency in AI

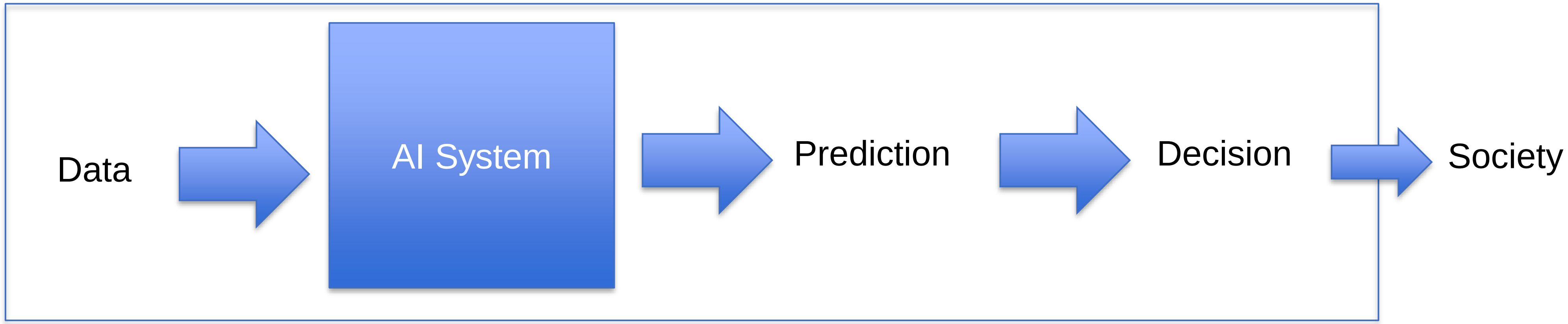
Transparency is a property of a system that makes it possible to get certain information regarding a system's inner workings



Transparency itself is ethically neutral and is not an ethical concept

Transparency is something that can manifest in many different ways, and something that can present a solution for underlying ethical questions.

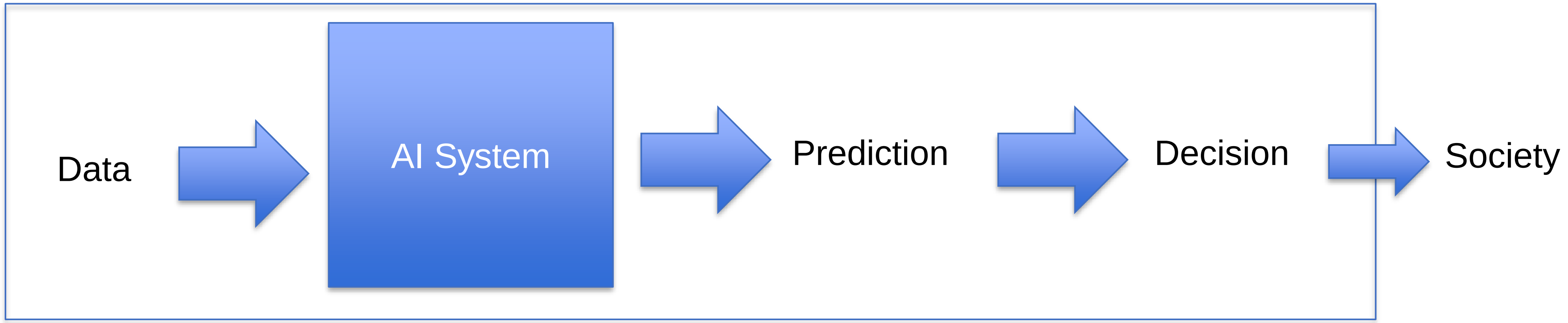
Transparency in AI - The justification of decisions



Transparency is relevant at least to the three following issues:

1. **The justification of decisions :**
   - Good governance in public or private sectors involves non-arbitrariness of decisions.
   - Applied to any kind of decision-making that has an ethically or legally relevant effect on individuals
   - **Non-arbitrariness** means access to justifications about "**why was this decision reached, and on what grounds?**"
   - Case of public governance, the capacity to contest and appeal are crucial. This represents a demand to right wrongs.

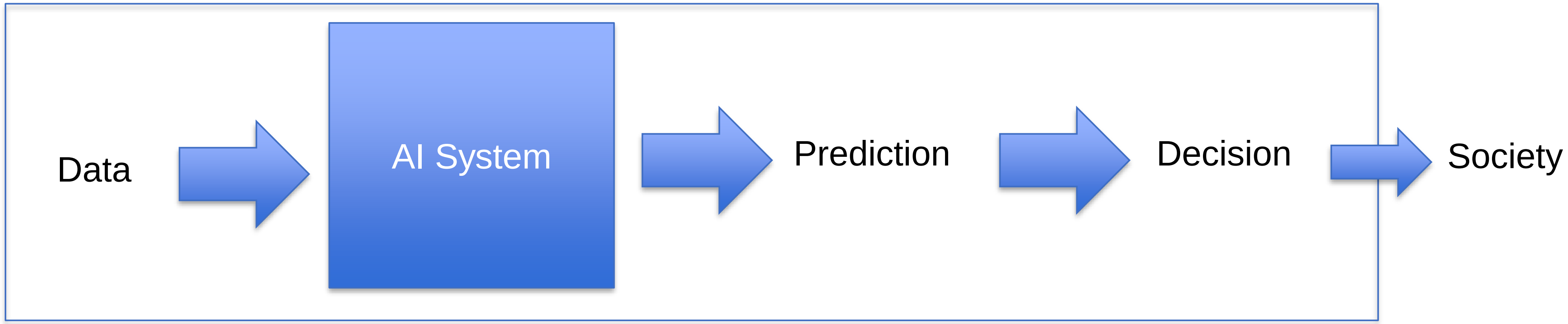
Transparency in AI – A Right to Know



Transparency is relevant at least to the three following issues:

2. **Right to Know :**
   - According to human rights, people are entitled to have explanations on how decisions were made so that they can maintain genuine agency, freedom and privacy
   - Freedom entails the right to get answers to questions such as **"How am I being tracked**? **What kind of inferences are being made about me**? **And how, exactly, have the inferences about me been made?"**

Transparency in AI – Moral Obligation

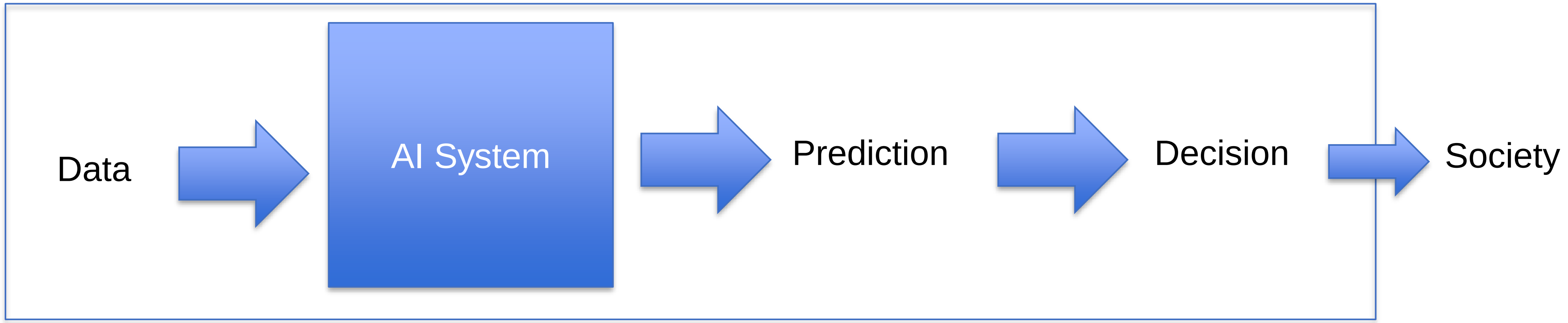


Transparency is relevant at least to the three following issues:

3. moral obligation to understand the consequences of our actions:

- Moral obligation, up to some reasonable level, to understand and predict the consequences of the kinds of technologies one brings into the world
- Stating "we can't understand now what it will do" is not a valid argument for unleashing a system that causes harm. Instead, it is our moral duty to explore the possible risks.

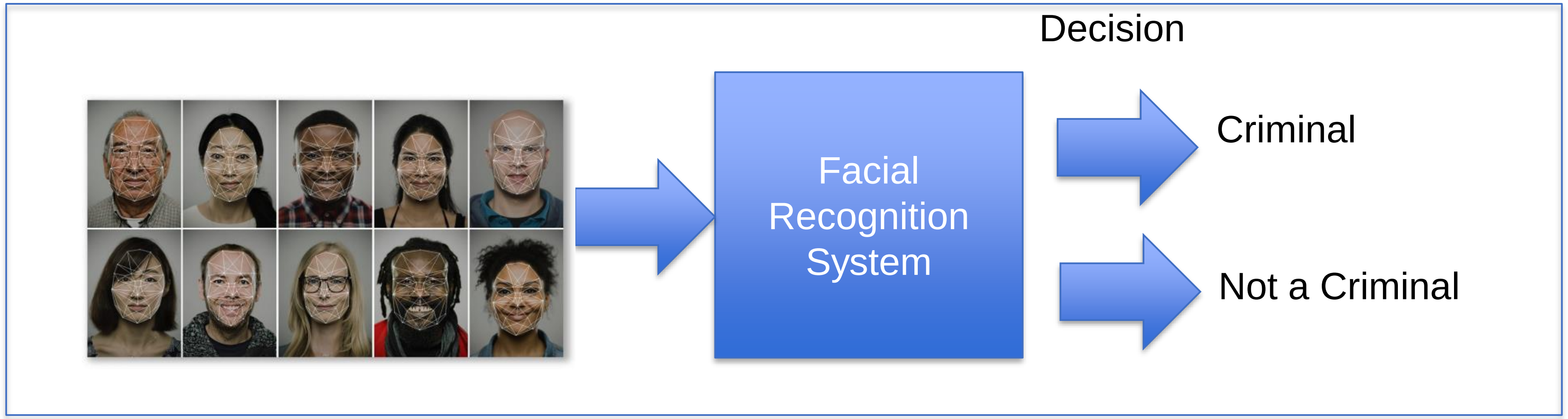
Transparency in AI – Summary



Transparency is relevant for call to <u>sufficient Information</u>

- Do we know whether and to what extent this **algorithmic decision is justified**?
- Do I know **how inferences about me are** made?
- To what extent I am **responsible for the actions of the system**
- How much I should **know about the inner workings** of the system to be able to take that responsibility?

Transparency in AI – Face Recognition



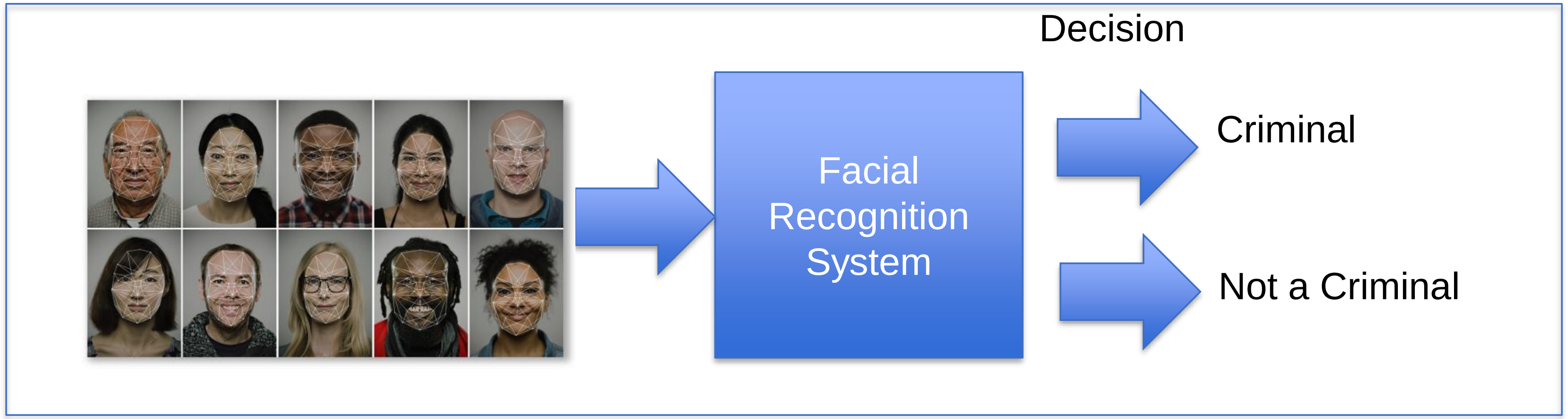
Face Recognition system used for security in Airport

- Systems starts **mis-categorizing individuals** who are not criminals as criminals
- Result – **Several innocent people are arrested**

- Transparency Lens:
  - Why did the system made mistakes
  - Explain why it made mistakes
  - Why should it matter

https://www.kaspersky.com/resource-center/definitions/what-is-facial-recognition

Transparency in AI – Challenges



Face Recognition system used for security in Airport

- Some contemporary machine learning systems are **so-called "black box" systems**, meaning we can't really see how they work. **This "opacity", or lack of visibility**, can be a problem if we use these systems to make decisions that have an effect on individuals.

- **Individuals have a right to know how critical decisions** – such as who gets accepted for a loan application, who gets paroled, and who gets hired – are made. This has led many to call for "more transparent AI".

Professor Note – Black box vs. White box AI models

Transparency in AI – Challenges – Black Box Models

https://www.brookings.edu/research/who-thought-it-was-a-good-idea-to-have-facial-recognition-software/

Industry Viewpoint

How China is building an all-seeing surveillance state

https://www.youtube.com/watch?v=uReVvICTrCM

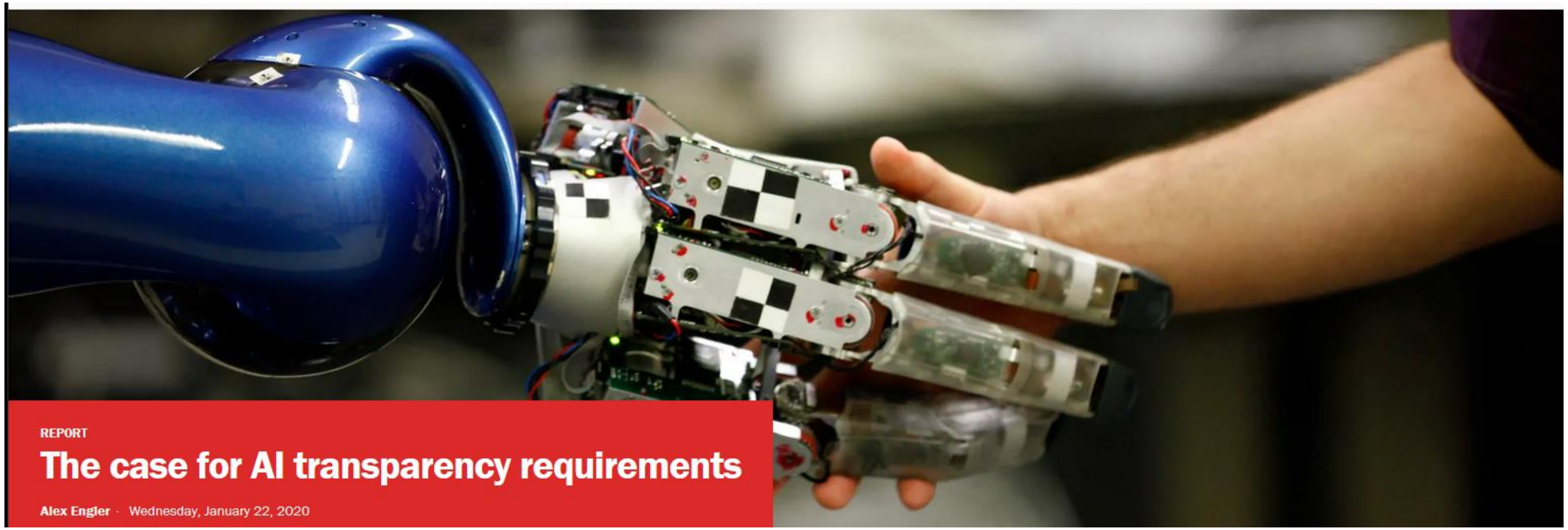
Transparency in AI – Challenges – Black Box Models

https://www.brookings.edu/research/10-actions-that-will-protect-people-from-facial-recognition-software/

Transparency in AI – Challenges – Black Box Models

https://www.brookings.edu/research/enrollment-algorithms-are-contributing-to-the-crises-of-higher-education/

Transparency in AI – Challenges – Black Box Models



https://www.brookings.edu/research/the-case-for-ai-transparency-requirements/

Transparency in AI – Challenges – Black Box Models

https://www.brookings.edu/research/how-to-improve-technical-expertise-for-judges-in-ai-related-litigation/

Transparency in AI – Challenges – Black Box Models

CITE [#]

SOCIAL

DOWNLOAD

CONTENTS

Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From an Explainable AI Competition

by Cynthia Rudin and Joanna Radin

Published on Nov 22, 2019

https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8

Transparency in AI – Challenges – Black Box Models

TECH / TRANSPORTATION / CARS

Two new fatal Tesla crashes are being examined by US investigators

Photo by James Bareham / The Verge

/ A pedestrian was killed in California, and two other people were killed in Florida

By ANDREW J. HAWKINS / @andyjayhawk

Jul 7, 2022, 4:16 PM EDT | 0 Comments / 0 New

https://www.theverge.com/2022/7/7/23198997/tesla-fatal-crashes-california-florida-autopilot-nhtsa

Industry Viewpoint

Building Trust In AI: The Case For Transparency

https://www.youtube.com/watch?v=nXtwXOoBcRw

What is Transparency

Transparency is, roughly, a property of an application

How much it is possible to understand about a system’s inner workings “in theory”

The way of providing explanations of algorithmic models and decisions that are comprehensible for the user.

Public perception and understanding of how AI works. Transparency can also be taken as a broader socio-technical and normative ideal of “openness

There are many open questions regarding what constitutes transparency or explain ability, and what level of transparency is sufficient for different stakeholders. Depending on the specific situation, the precise meaning of “transparency” may vary.

It is an open scientific question, whether there are several different kinds, or types, of transparency - analyze the legal significance of unjust biases or to discuss them in terms of features of machine learning systems.

What is Transparency- Our FOCUS

Transparency as a property of a system

“Explainability” (AI research in this area is known as “XAI”), “interpretability”, “understandability”, and “black box

Transparency as a property of a system

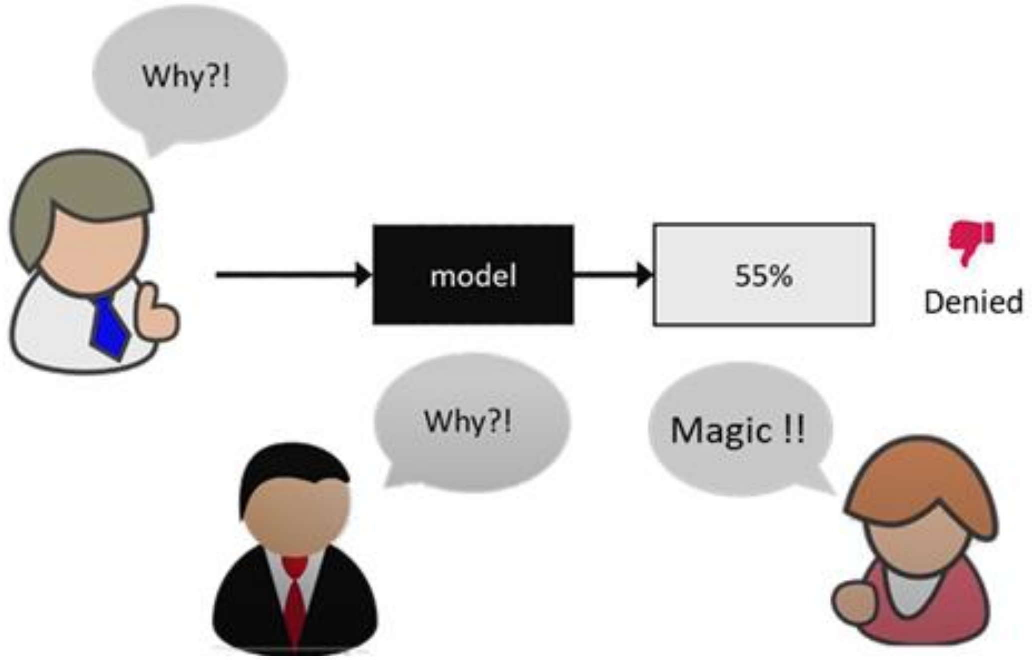
As a property of a system, transparency addresses how a model works or functions internally

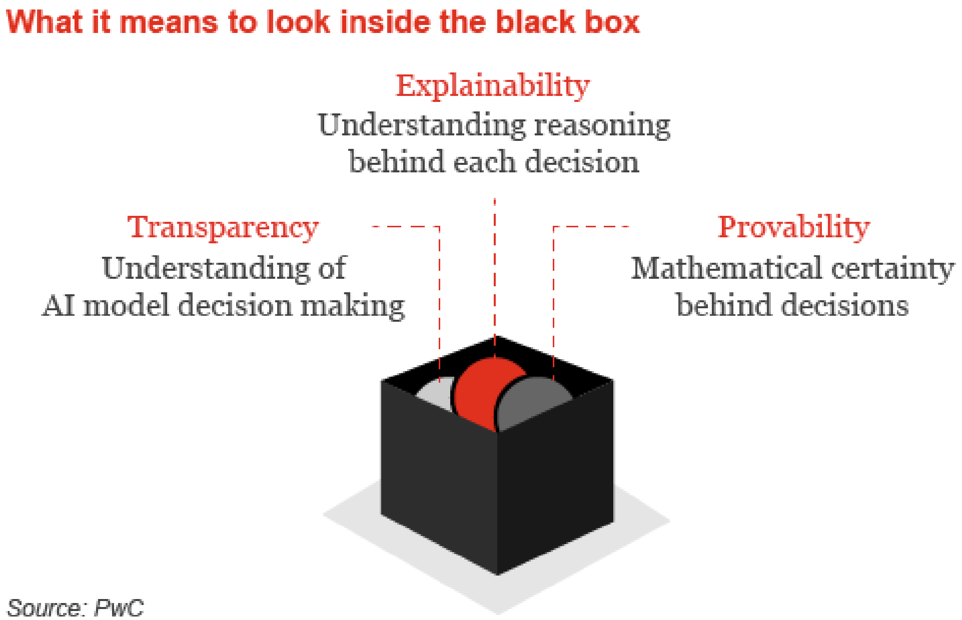
Divided into:

1. **Simulatability – <u>An understanding of the functioning of the model</u>**
2. **Decomposability – Understanding the individual components**
3. **Algorithm transparency – Visibility of algorithms**

Module 3 – XAI

What is a Black BOX System- Explainability

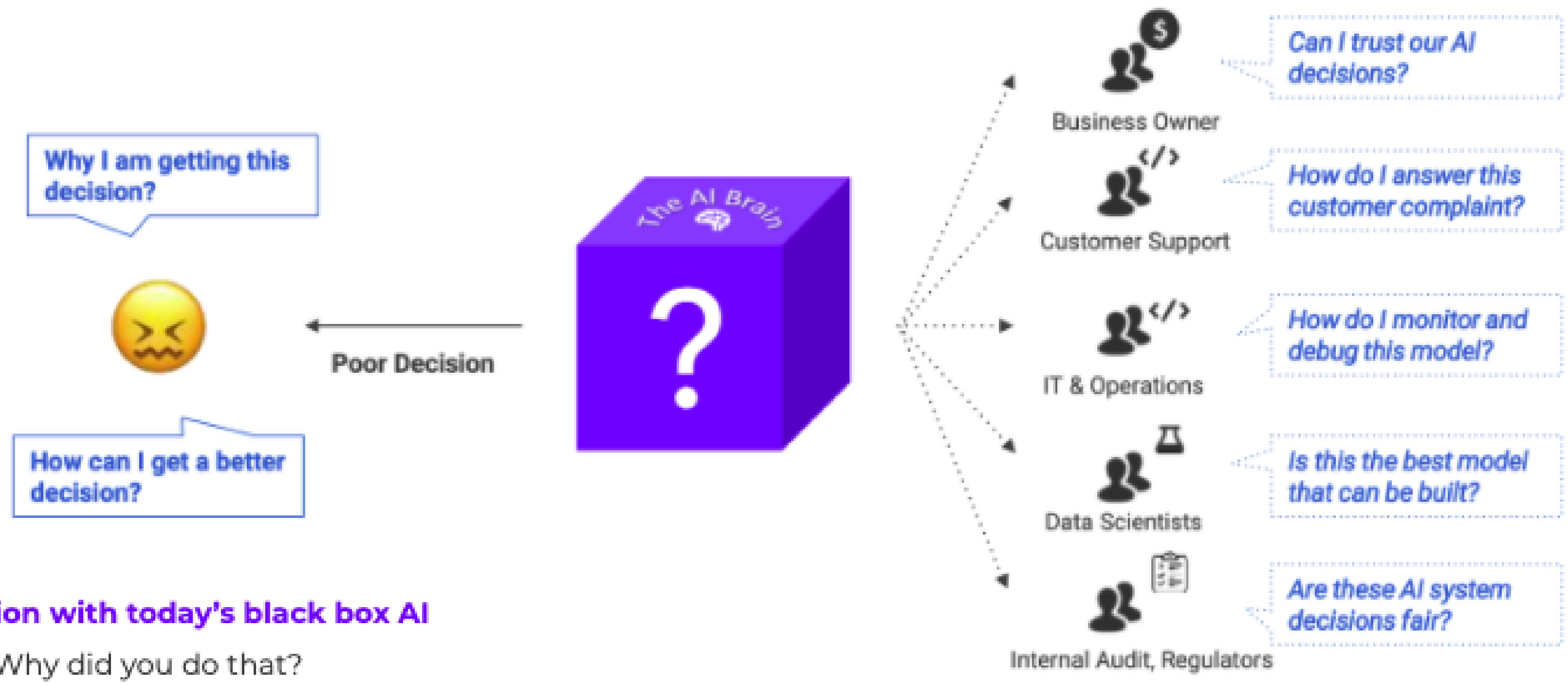




Source: PwC

Black BOX vs. White BOX Algos

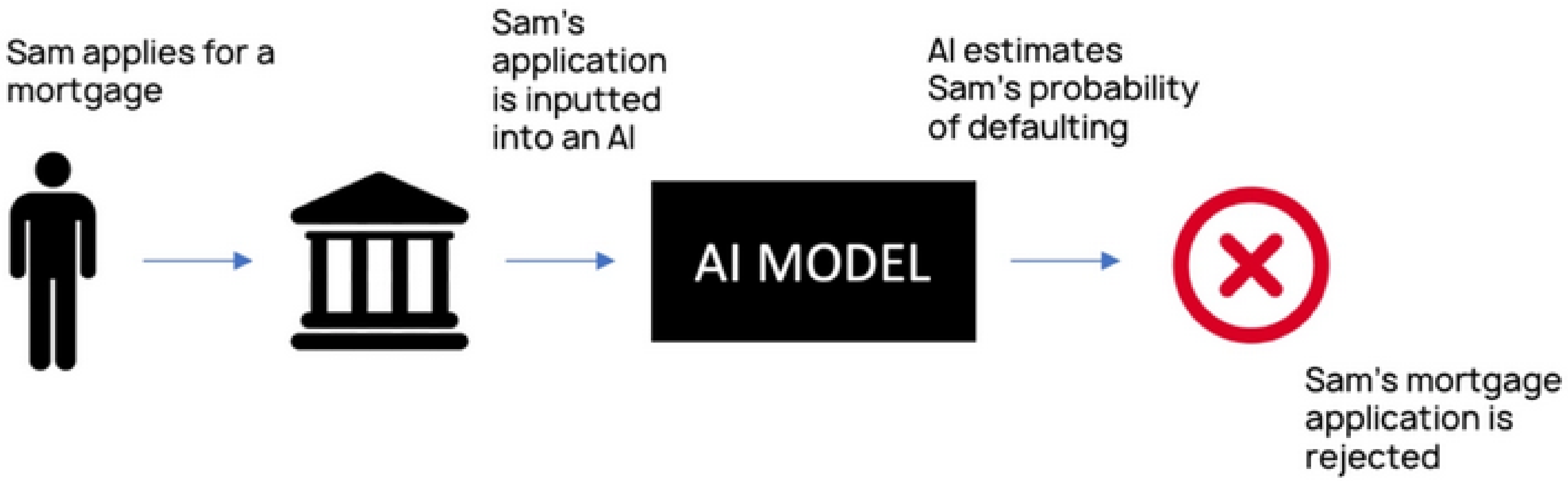
Black-box AI creates confusion & doubt



Confusion with today's black box AI

- Why did you do that?
- Why did you not do that?
- When do you succeed or fail?
- How do I correct an error?

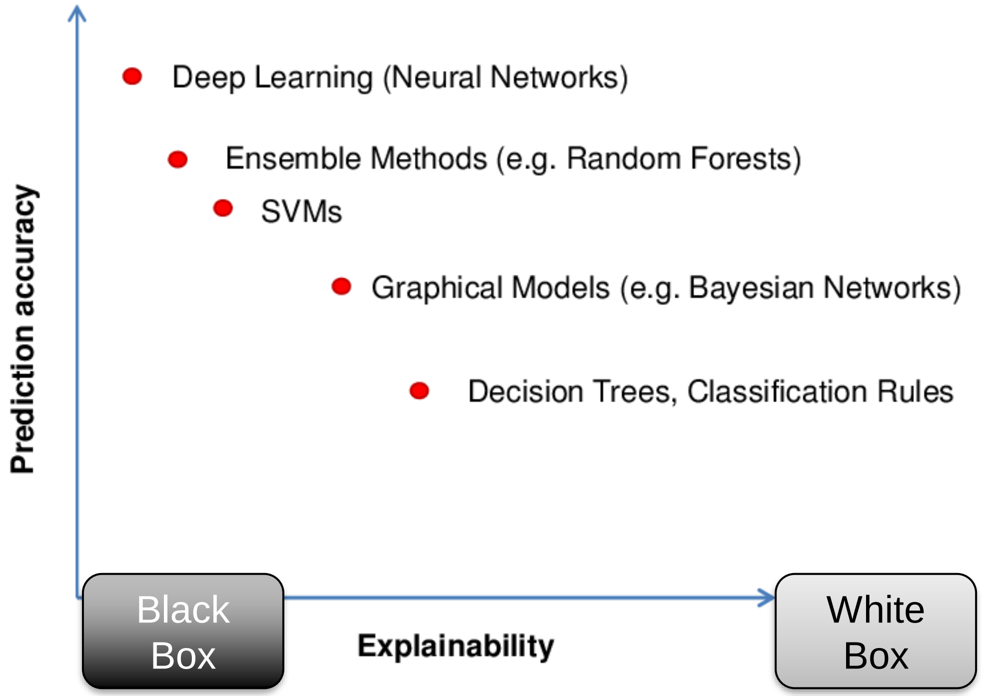
Black BOX vs. White BOX Algos

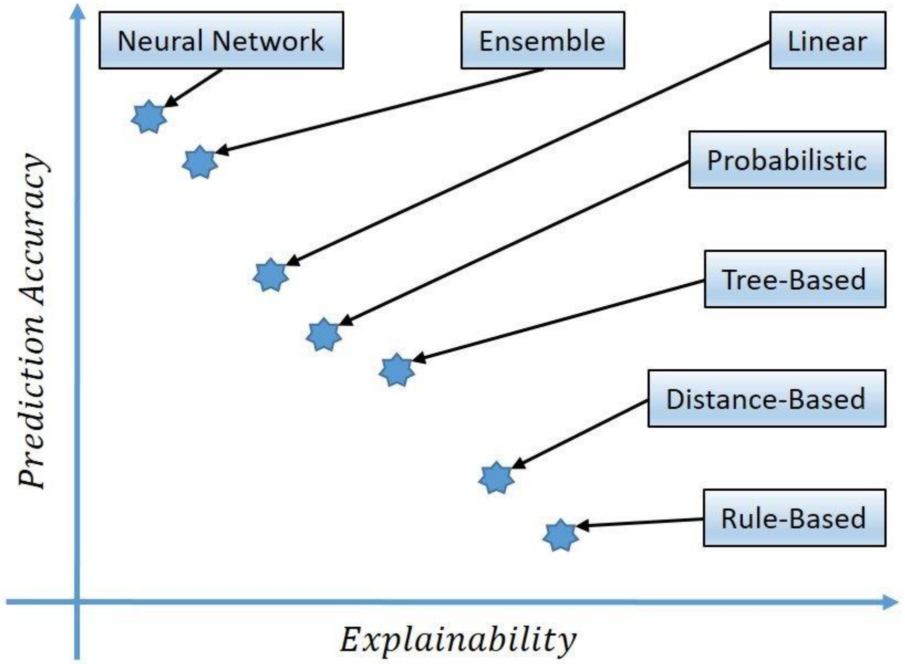


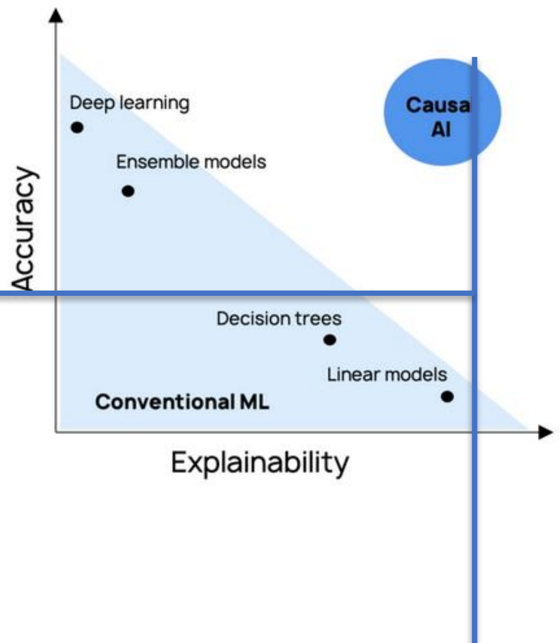
A simple example of an AI use case: an AI model decides which mortgages to approve.

https://www.causalens.com/blog/xai-doesnt-explain/

Black BOX vs. White BOX Algos – Why are using Black BOX everywhere ????

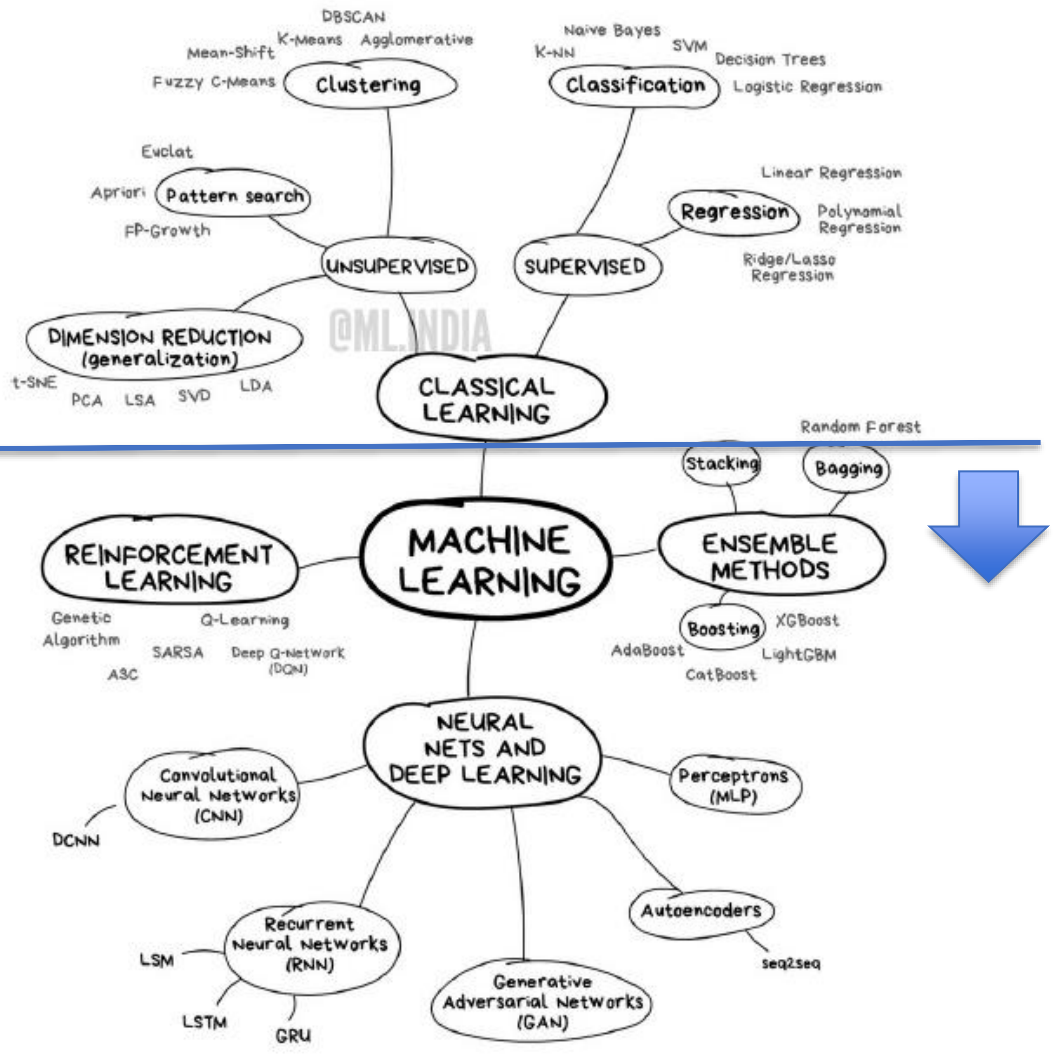






Algo Map – Choose Carefully

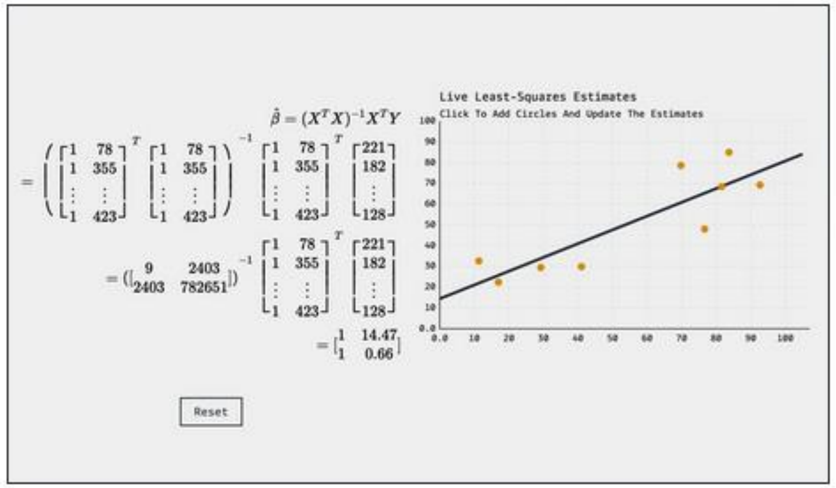
ML ALGO MAP



Do not

Image by vas3k.

White BOX Algo – Demo in Class

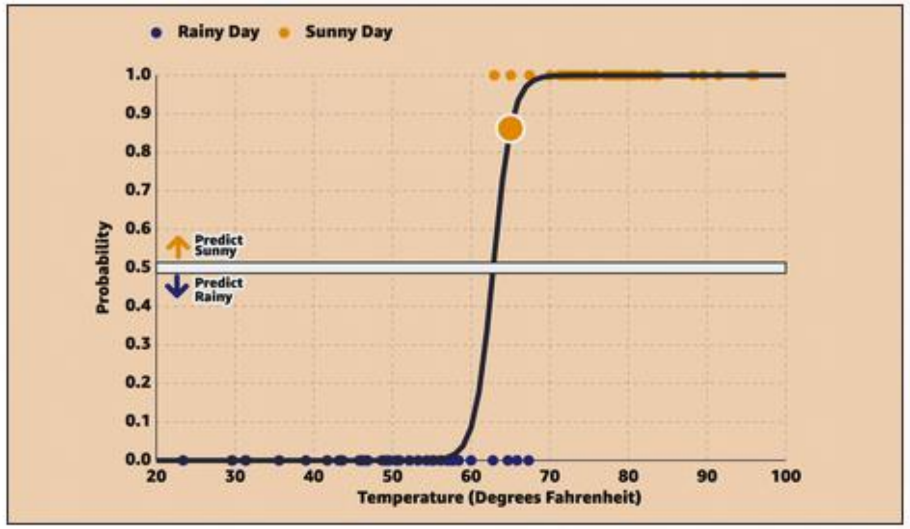


LINEAR REGRESSION

Interactively learn about linear regression models as they're commonly used in the context of machine learning.

Dive In

https://mlu-explain.github.io/linear-regression/



LOGISTIC REGRESSION

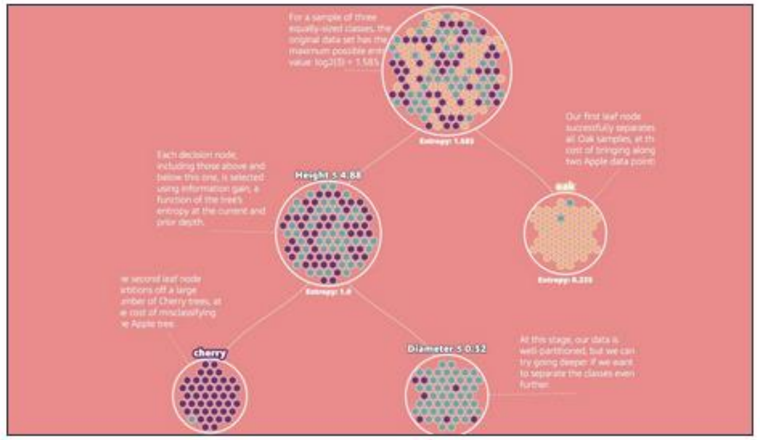
Learn how logistic regression can be used for binary classification in machine learning through an interactive example.

Dive In

https://mlu-explain.github.io/logistic-regression/

Homework – Summarize

White BOX Algo



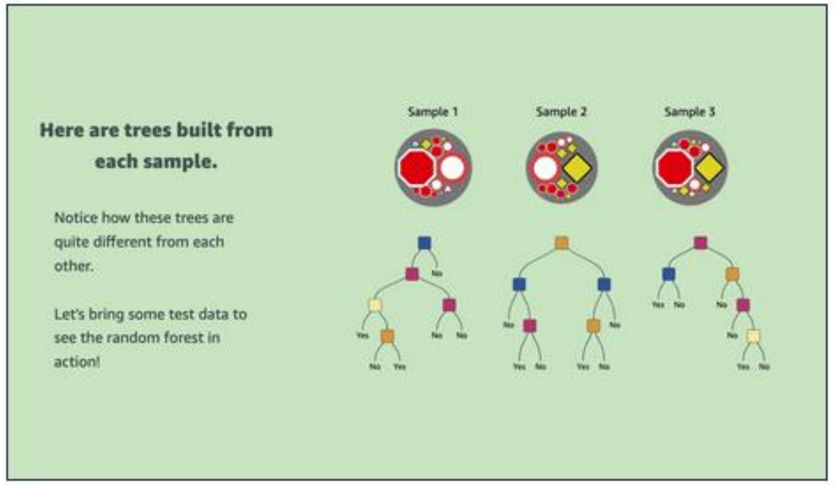
DECISION TREES

Explore one of machine learning's most popular supervised algorithms: the Decision Tree. Learn how the tree makes its splits, the concepts of Entropy and Information Gain, and why going too deep is problematic.

Dive In

https://mlu-explain.github.io/decision-tree/

Black BOX Algo



RANDOM FOREST

Learn how the majority vote and well-placed randomness can extend the decision tree model to one of machine learning's most widely-used algorithms, the Random Forest.

Dive In

https://mlu-explain.github.io/random-forest/

Homework – Walk through in class

Industry Viewpoint

Explainable AI explained!

https://www.youtube.com/watch?v=OZJ1IgSgP9E

COMPAS Recidivism Algorithm

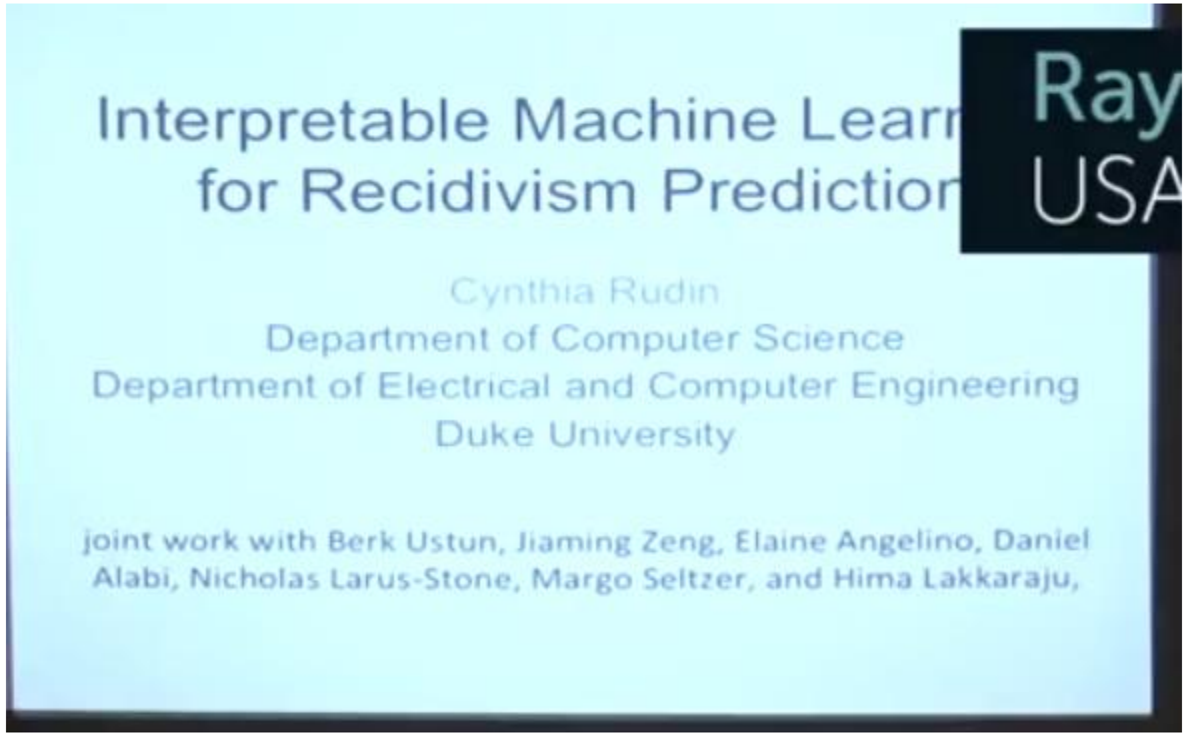
How We Analyzed the COMPAS Recidivism Algorithm

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

May 23, 2016

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Cynthia Rudin - Interpretable ML for Recidivism Prediction - The Frontiers of Machine Learning

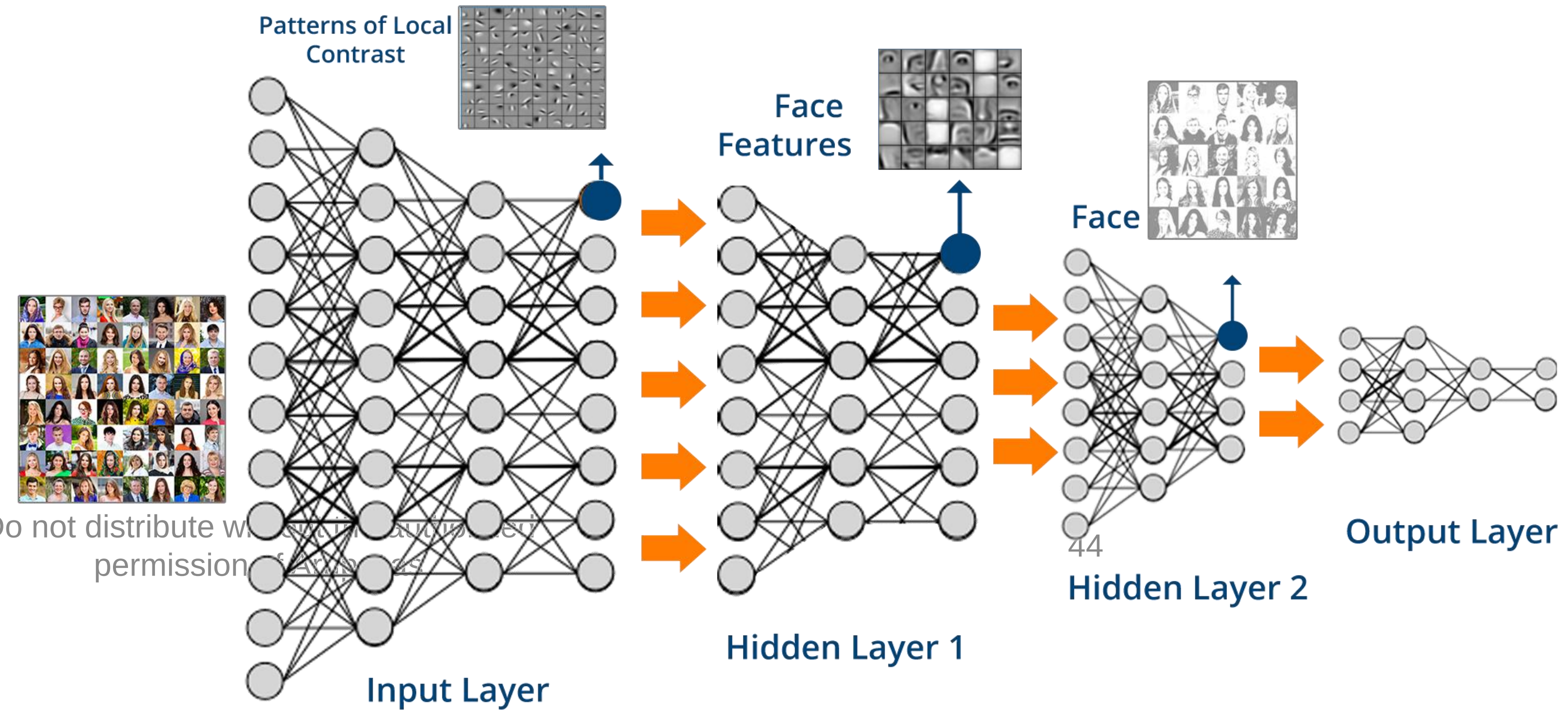


https://www.youtube.com/watch?v=MjxcwKN2dXs

What makes a system a Black Box - Complexity

Complexity:

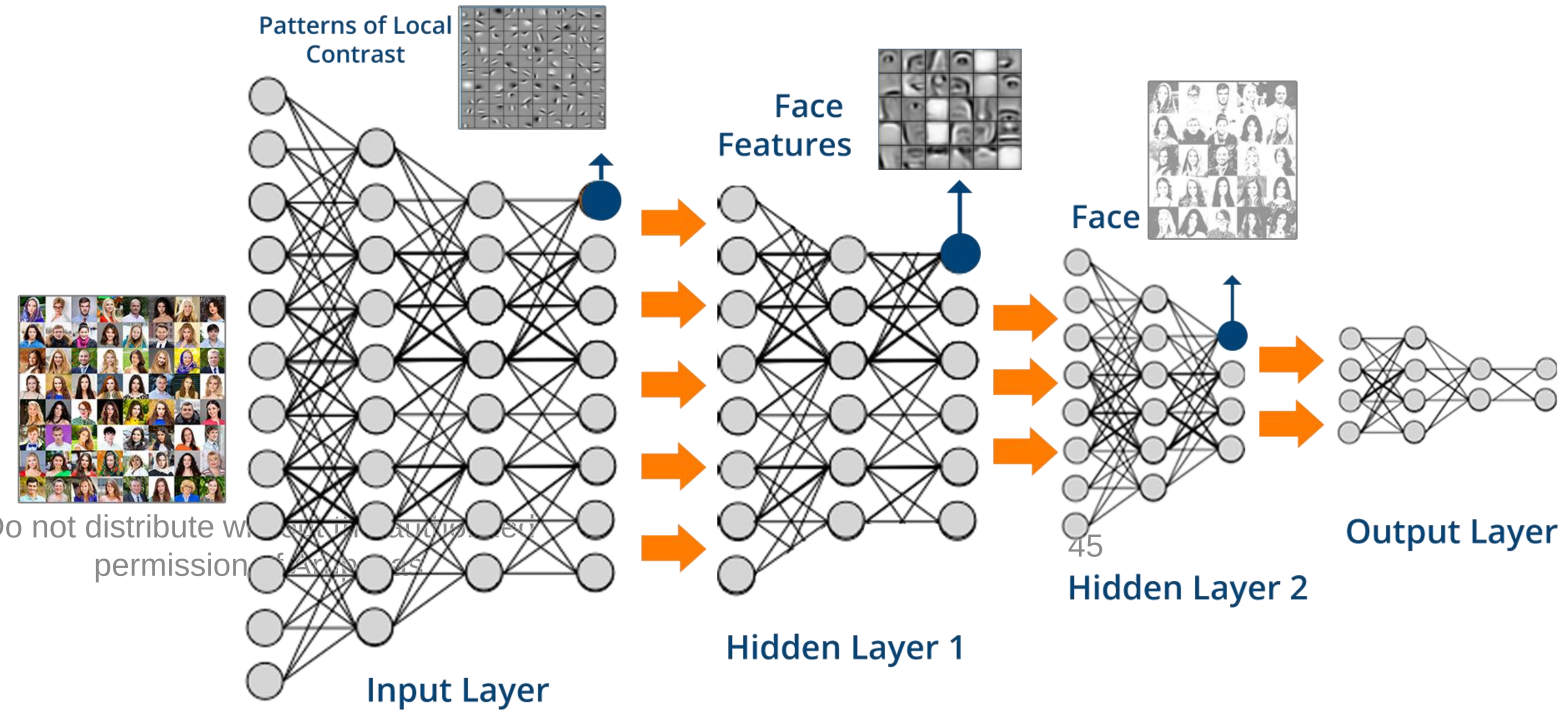
- In contemporary AI-systems, operation of a neural network is encoded in **thousands, or <u>even millions, of numerical coefficients</u>**.
- Typically the system learns their values at the training phase.
- Because the operation of the neural network **depends on the complicated interactions between these values, it is practically impossible to understand how the network works even if all the parameters are known**.



What makes a system a Black Box – Difficulty in Explainability

Difficulty of developing explainable solutions.

- Even if the used AI models support some level of explainability
- additional development is required to build explainability to the system
- It may be **difficult to create a user experience for careful yet easily understandable explanations for the users.**



Black BOX models - Explainability

Majority of AI systems today are using Deep learning Black BOX models ---- Why ????

Impossible to get full transparency as the mathematical model are complicated involving million of terms

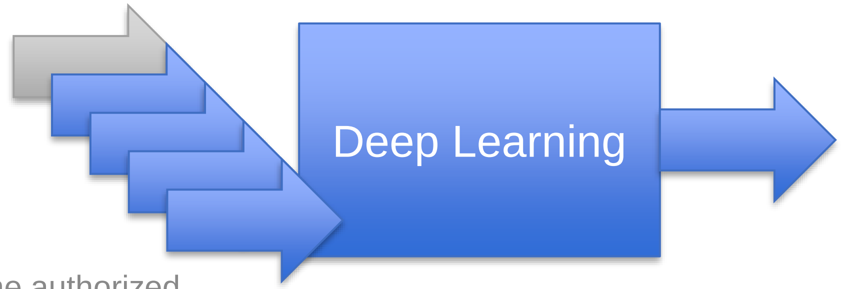
Comprise – Sufficient level of transparency *(Would it suffice if algorithms offered people a disclosure of how algorithms came to their decision and* ***provide the smallest change "that can be made to obtain a desirable outcome****" (Wachter et al., 2018)*

For example, if an algorithm refuses someone a social benefit, it should tell the person the reason, and also what he or she can do to reverse the decision.

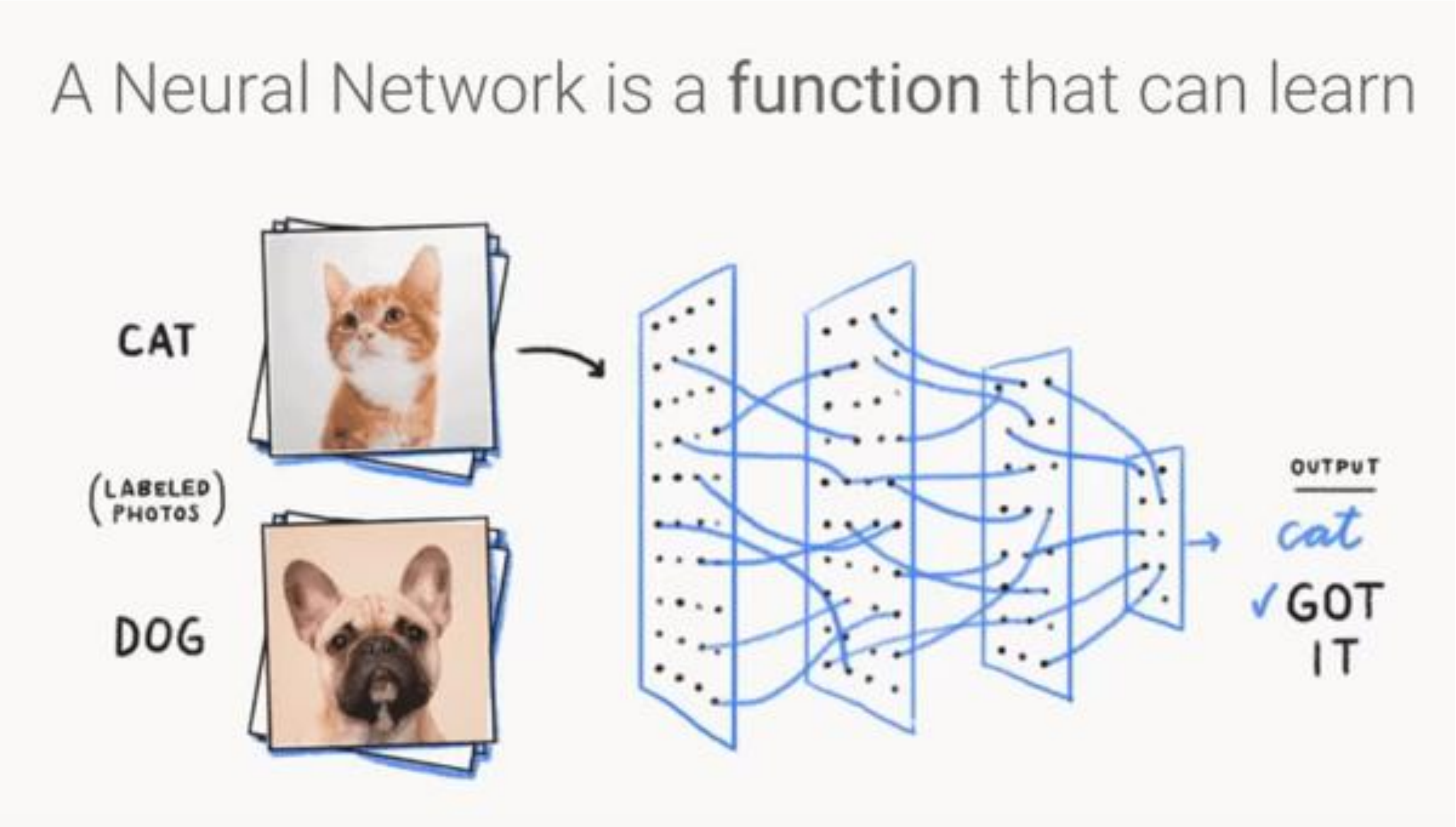
Black BOX models - Explainability

- The **explanation should tell, for instance, what the maximum amount of salary to be approved is (input), and how decreasing the amount will impact the decisions made** (manipulation of the input)

- But the problem is that the right to know also applies to situations where the system makes mistakes. Then, it may be necessary to perform an autopsy on the algorithm and identify those factors that caused the system to make mistakes (Rusanen & Ylikoski 2017). This can’t be done by only manipulating the inputs and outputs.



Black BOX models - Explainability



The model has inferred two patterns that make up a cat. To the model, they're just numbers, but to us, they look like describable patterns

https://distill.pub/2017/feature-visualization/

Transparency- Comprehensibility

- The comprehensibility – or understandability – of an algorithm requires that one should explain how a decision was made by an AI model in a way that is sufficiently understandable to those affected by the model. One should have a concrete sense of how or why a particular decision has been arrived at based on inputs.

- **Difficult to translate algorithmically derived concepts into human-understandable concepts**. In some countries, legislators have discussed whether public authorities should publish the algorithms they use in automated decision-making in terms of programming codes. However, most people do not know how to make sense of programming codes**. It is thus hard to see how transparency is increased by publishing codes.**

```
#Part 2 - Fitting the CNN to the images
from keras.preprocessing.image import ImageDataGenerator

train_datagen = ImageDataGenerator(
        rescale=1./255,
        shear_range=0.2,
        zoom_range=0.2,
        horizontal_flip=True)

test_datagen = ImageDataGenerator(rescale=1./255)

training_set = train_datagen.flow_from_directory(
        'dataset/training_set',
        target_size=(64, 64),
        batch_size=32,
        class_mode='binary')

test_set= test_datagen.flow_from_directory(
        'dataset/test_set',
        target_size=(64, 64),
        batch_size=32,
        class_mode='binary')
```

Transparency- AI Model Cards – A possible solution

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data

- CelebA [36], training data split.

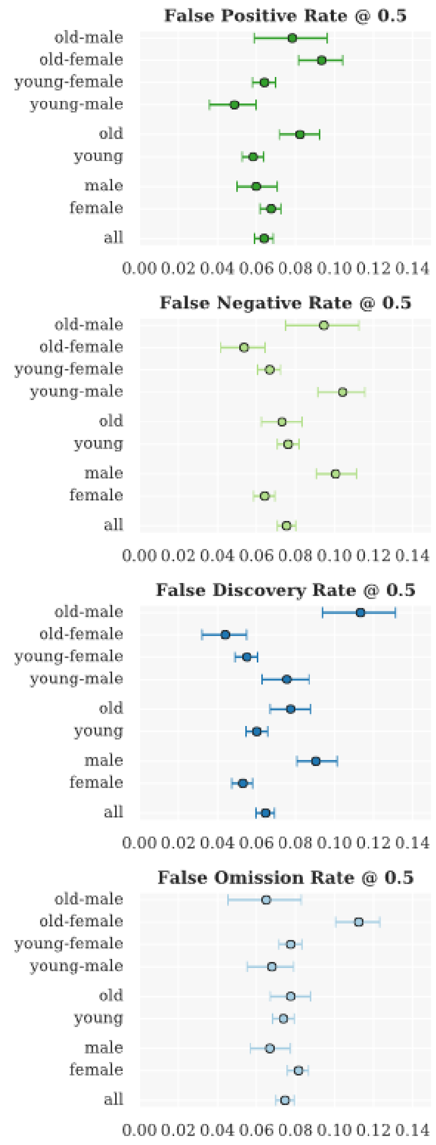
Evaluation Data

- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Quantitative Analyses

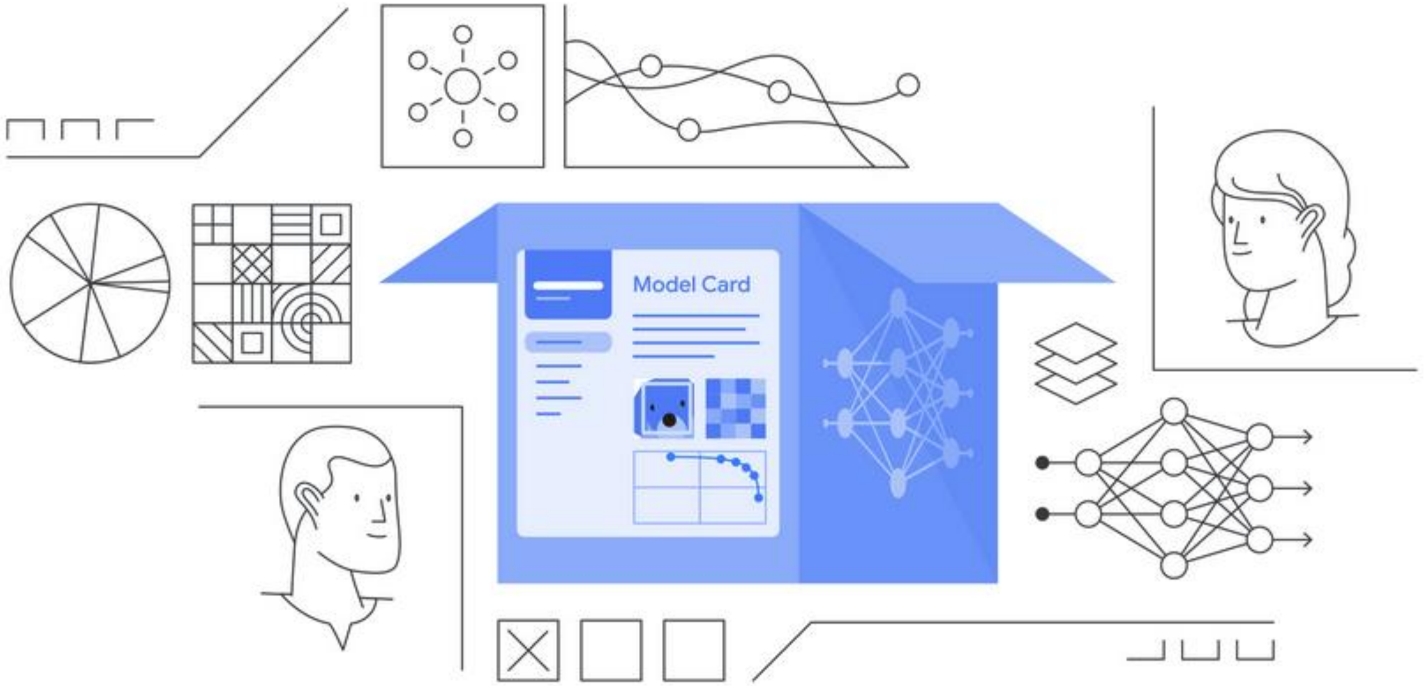


Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Transparency- AI Model Cards – A possible solution



Homework – Summarize

https://modelcards.withgoogle.com/about

Transparency- AI Model Cards – A possible solution

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

https://arxiv.org/pdf/1810.03993.pdf

How to make models more transparent?

The black box problem of artificial intelligence is not new. Providing transparency for machine learning models is an active area of research. Roughly speaking, there are five main approaches:

- **Use simpler models**. This, however, often sacrifices accuracy for explainability.
- **Combine simpler and more sophisticated models**. While the sophisticated model allows the system to do more complex computations, the simpler model can be used to provide transparency.
- **Modify inputs to track relevant dependencies between inputs and outputs**. If a manipulation of inputs changes overall model results, these inputs may play a role in the classification.
- Design the models for the user. This requires using cognitively and psychologically efficient methods and tools for visualizing the model states or directing attention. For example, in computer vision, states in intermediate layers of the models can be visualized as features (like heads, arms, and legs) to provide a comprehensible description for image classification. Researchers have also developed methods for directing “attention” towards the parts of the input that matter the most. These can be visualized to highlight the parts of an image or a text (so-called “weights”) that contribute the most to a particular recommendation.
- **Follow the latest research**. A lot of research is ongoing on various aspects of explainable AI – including the socio-cognitive dimensions – and new techniques are being developed.

Module 3 – Transparency and the risks of openness

Transparency and the risks of openness

Transparency often denotes a modern, ethico-socio-legal “ideal” (Koivisto 2016), a normative demand for the acceptable use of technology in our societies.

Paradoxically, the ideal of openness can lean to harmful consequences, too.

- For example, **the transparency of social media platforms has led to several instances of misuse and democratic challenge**s.
- Transparency can create security risks.
- Too much transparency may lead to leaking of privacy-sensitive data into the wrong hands. Or the more that is revealed about the algorithms and the data, the more harm a malicious actor can cause.
- **Algorithms can be hacked**, and information may make AI more vulnerable to intentional attacks.
- **Entire algorithms can also be stolen based simply on their explanations alone**.

Module 3 – Summary

Summary

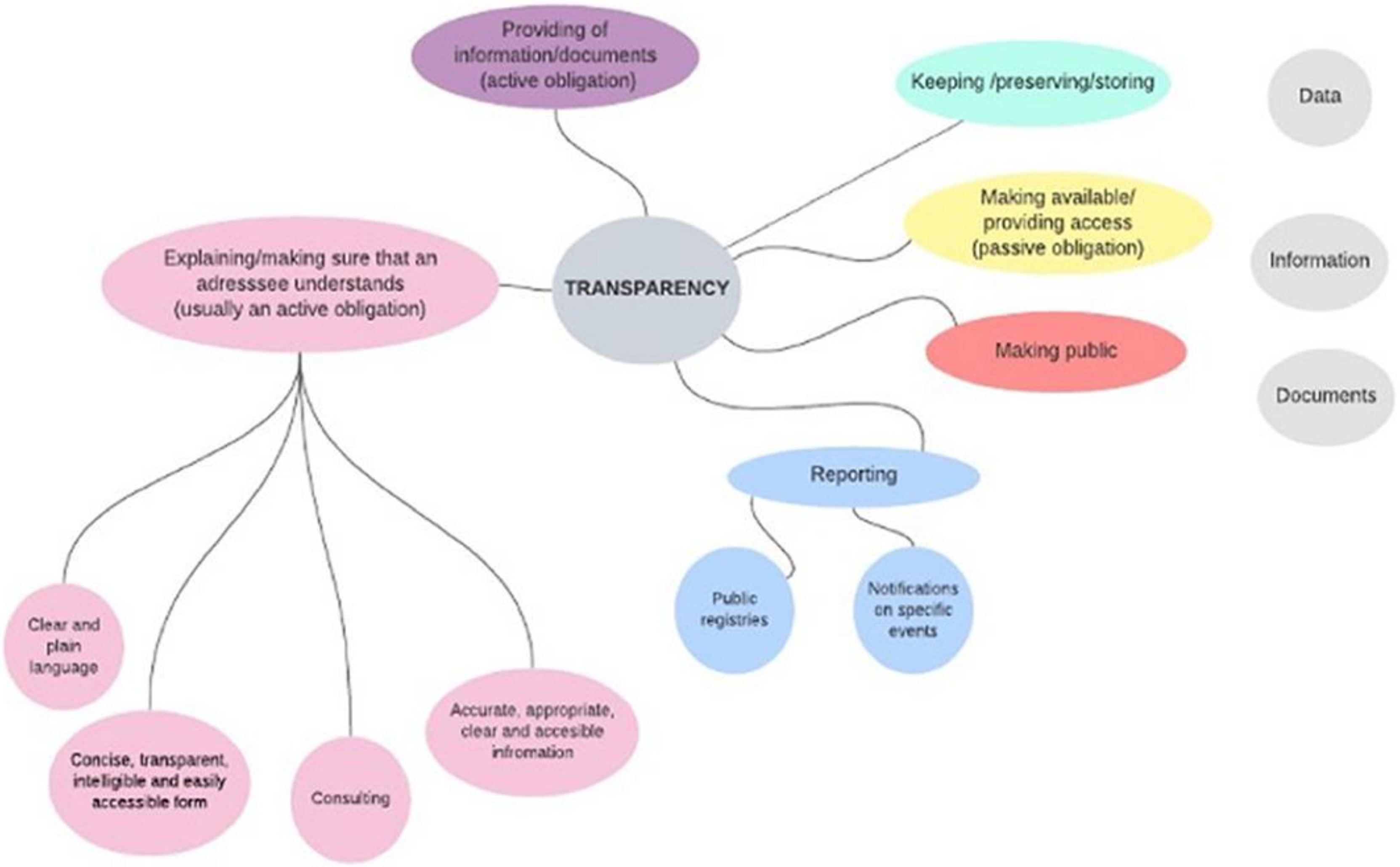
While there is a need to develop more transparent practices for AI, there is also a need to develop practices that can help us to avoid abuse.

While transparency may help to mitigate ethical issues – such as fairness or accountability – it also creates ethically important risks.

Too much openness in the wrong context may defeat the positive development of AI-enabled processes.

Taken together, it is clear that the ideal of full transparency of algorithms should be carefully considered, and we will have to **<u>find a balance between security and transparency considerations.</u>**

Summary



https://www.frontiersin.org/articles/10.3389/frai.2022.879603/full

Module 3: Additional Readings

The AI Transparency Paradox

[No Title]

by Andrew Burt

December 13, 2019

Jorg Greuel/Getty Images

Summary. In recent years, academics and practitioners alike have called for greater transparency into the inner workings of artificial intelligence models, and for many good reasons. Transparency can help mitigate issues of fairness, discrimination, and trust — all of which have received increased attention. At the same time, however, it is becoming clear that disclosures about AI pose their own risks: Explanations can be hacked, releasing additional information may make AI more vulnerable to attacks, and disclosures can make companies more susceptible to lawsuits or regulatory action. Call it AI's "transparency paradox" — while generating more information about AI might create real benefits, it may also lead to new downsides. To navigate this paradox, organizations will need to think carefully about how they're managing the risks of AI, the information they're generating about these risks, and how that information is shared and protected. **close**

https://hbr.org/2019/12/the-ai-transparency-paradox

Business Ethics

Building Transparency into AI Projects

by Reid Blackman and Beena Ammanath

June 20, 2022

Illustration: Nata Schepy

Summary. As algorithms and AIs become ever more embedded in people's lives, there's also a growing demand for transparency around when an AI is used and what it's being used for. That means communicating why an AI solution was chosen, how it was designed and developed, on what grounds it was deployed, how it's monitored and updated, and the conditions under which it may be retired. There are four specific effects of building in transparency: 1) it decreases the risk of error and misuse, 2) it distributes responsibility, 3) it enables internal and external oversight, and 4) it expresses respect for people. Transparency is not an all-or-nothing proposition, however. Companies need to find the right balance with regards to how transparent to be with which stakeholders. **close**

https://hbr.org/2022/06/building-transparency-into-ai-projects

OECD.AI Policy Observatory

OECD.org Going Digital Toolkit | EN

Experts & blog AI Principles Policy areas Trends & data Countries About

Home > OECD AI Principles > Transparency and explainability (Principle 1.3)

Transparency and explainability (Principle 1.3)

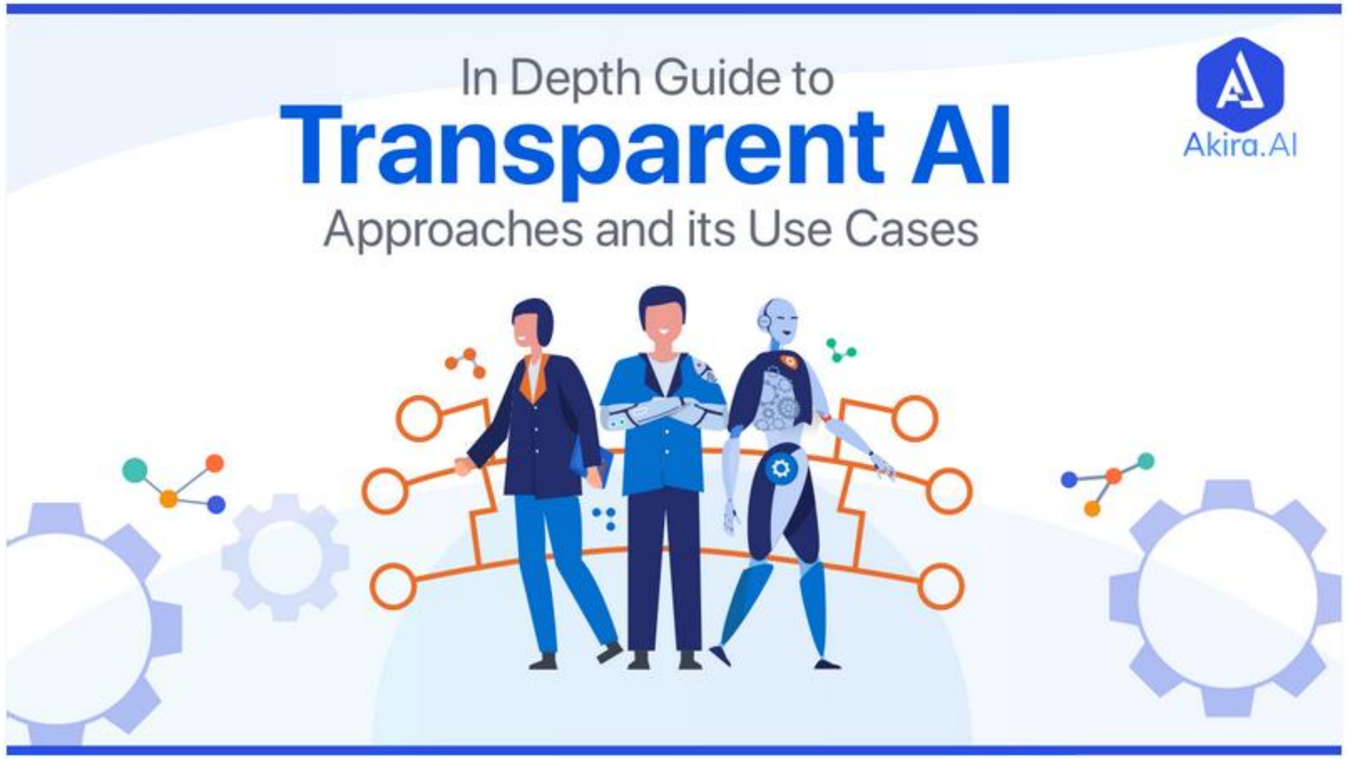
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

> *AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:*
>
> - *to foster a general understanding of AI systems,*
> - *to make stakeholders aware of their interactions with AI systems, including in the workplace,*
> - *to enable those affected by an AI system to understand the outcome, and,*
> - *to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.*

https://oecd.ai/en/dashboards/ai-principles/P7

Transparent AI Challenges and Its Solutions | Ultimate Guide

Jagreet Kaur Gill - posted on Nov 10, 2021 9:38:35 AM



https://www.akira.ai/blog/transparent-ai-challenges

Homework – Summarize

Accelerating the Path Towards AI Transparency

Aug 21,2020

IBM **Policy Lab**

Bold Ideas for a Digital Society

When you're grocery shopping, you may check an item's nutrition label to discover its calorie or sugar content. This crucial information helps people make informed decisions about their eating habits and ultimately their health. A similar kind of transparency is what we should expect in AI systems, especially when they are used in the context of high-stakes decisions, such as in healthcare, public or financial services, and justice.

https://www.ibm.com/policy/wp-content/uploads/2020/07/IBMPolicyLab-AI-Transparency-FactSheets.pdf

Francesca Rossi, IBM AI Ethics Global Leader and IBM Fellow

Aleksandra Mojsilović, IBM Research Head of AI Foundations, Co-Director of IBM Science for Social Good, and IBM Fellow

REVIEW article

Front. Artif. Intell., 30 May 2022
Sec. Medicine and Public Health
https://doi.org/10.3389/frai.2022.879603

This article is part of the Research Topic
Explainable Artificial Intelligence for Critical Healthcare Applications
View all Articles >

Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations

Anastasiya Kiseleva[1,2*], Dimitris Kotzinos[2†] and Paul De Hert[1†]

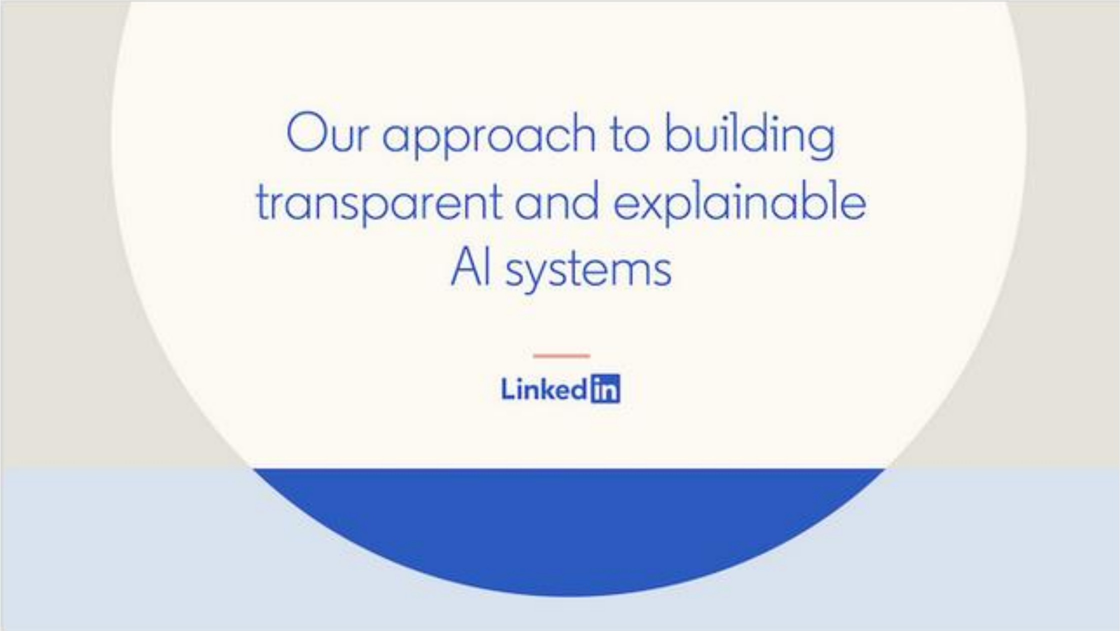
[1] LSTS Research Group (Law, Science, Technology and Society), Faculty of Law, Vrije Universiteit Brussels, Brussels, Belgium
[2] ETIS Research Lab, Faculty of Computer Science, CY Cergy Paris University, Cergy-Pontoise, France

https://www.frontiersin.org/articles/10.3389/frai.2022.879603/full

Our approach to building transparent and explainable AI systems

Kinjal Basu October 7, 2021

Share Tweet Share



https://engineering.linkedin.com/blog/2021/transparent-and-explainable-AI-systems

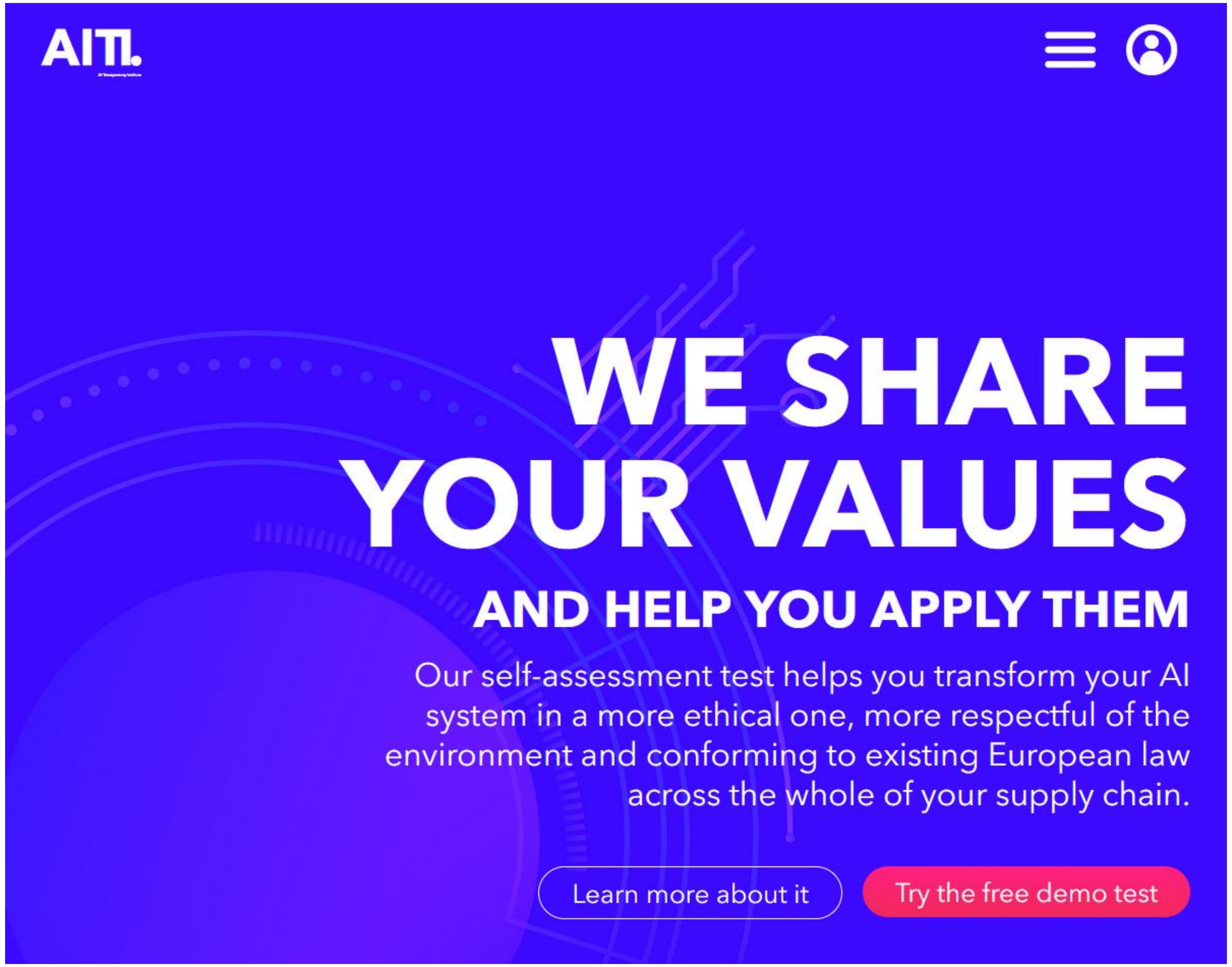
BROOKINGS

CLIMATE AI CITIES & REGIONS GLOBAL DEV INTL AFFAIRS U.S. ECONOMY U.S. POLITICS & GOVT MORE

AI GOVERNANCE

Artificial intelligence, machine learning, and data analytics are upending everything from education and transportation to health care and finance. In this series led by Governance Studies Vice President Darrell West, scholars from in and outside Brookings will identify key governance and norm issues related to AI and propose policy remedies to address the complex challenges associated with emerging technologies.

https://www.brookings.edu/series/ai-governance/



https://aitransparencyinstitute.com/

Appendix