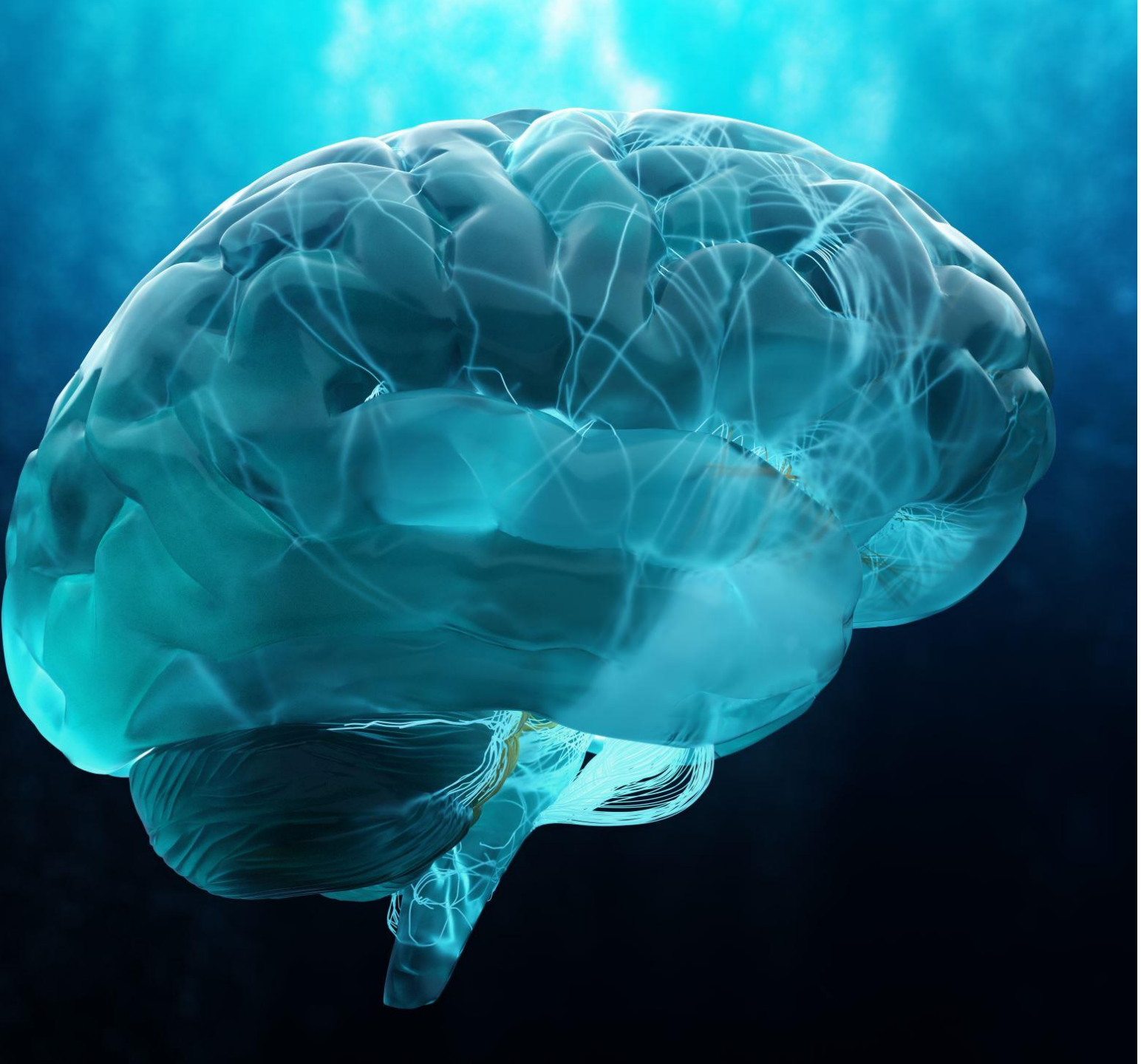
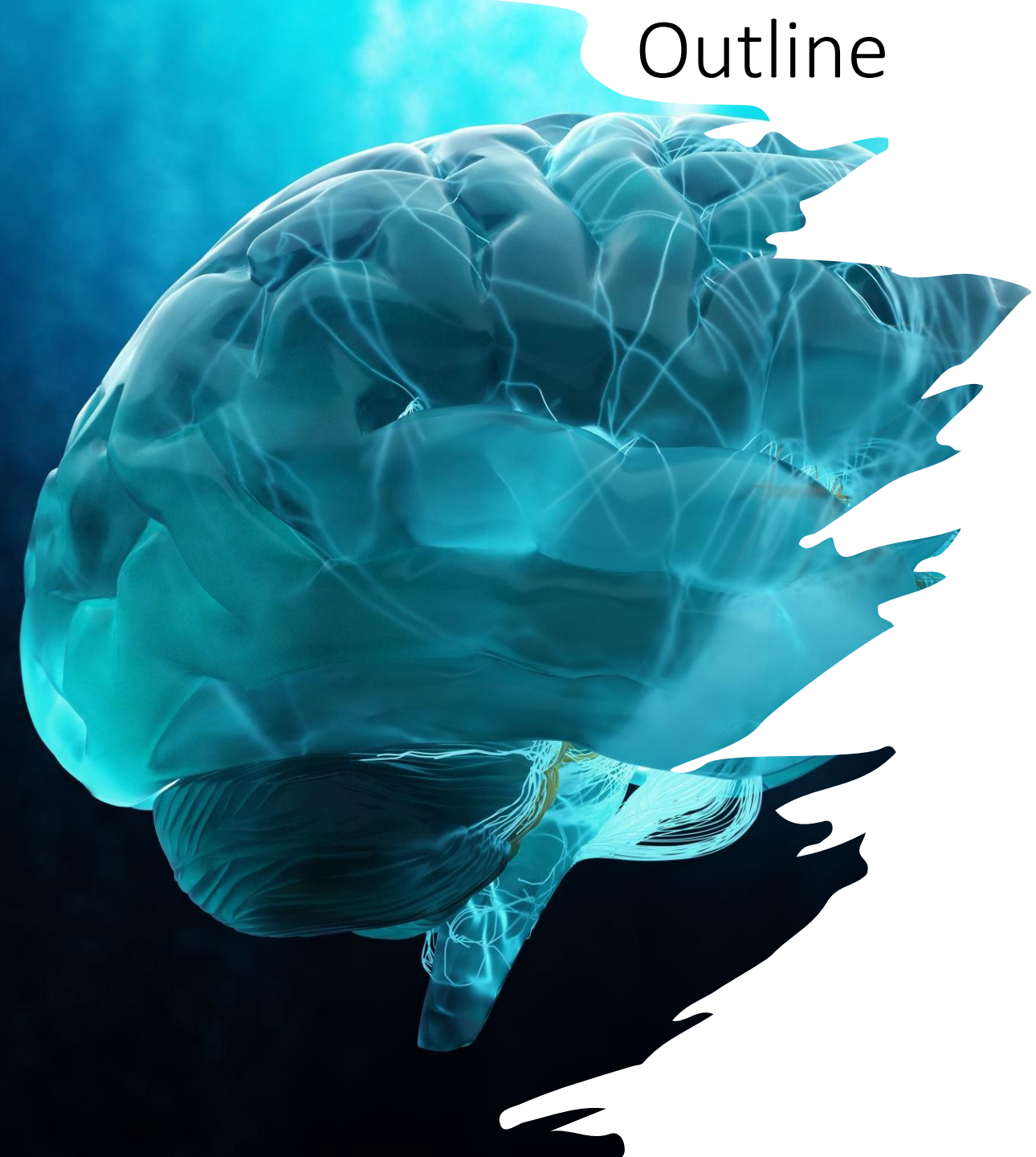


Managing Artificial Intelligence

Nahla Ben Amor
Professor in Business Computing
University of Tunis

nahla.benamor@gmx.fr
nah.benamor@gmail.com

July, 2023



Outline

- ▷ 1. Basics of Artificial Intelligence
  - ▷ 1.1 Definitions
    - Machine Learning
    - Deep Learning
  - ▷ 1.2 Main Domains of AI
- ▷ 2. Ethics and Responsibility in the Use of AI
  - ▷ 2.1 Ethical Issues Related to AI
  - ▷ 2.2 Sources of Bias in ML Lifecycle
  - ▷ 2.3 Regulatory and Normative Framework for AI
  - ▷ 2.4 Best Practices for Bias Avoidance/Mitigation
  - **Task at Hand: Ethical Implications and Trustworthy Use of AI**
- ▷ 3. Risk Management and Compliance with AI-Related Regulations
  - ▷ 3.1 Adversarial Machine Learning
  - ▷ 3.2 Risk Management in AI Systems
  - ▷ 3.3 AI Regulation
- ▷ 4. Impacts of AI on the World of Work
  - ▷ 4.1 Task Automation and Job Transformation
  - ▷ 4.2 How Artificial Intelligence Will Redefine Management?
- ▷ 5. Managing Artificial Intelligence
- ▷ Conclusion

Revision exercices of Part 1

Multiple-choice quiz: For each of the following questions select the correct answer(s):

1) What is the primary significance of Moore's Law?
a) It allows for a significant reduction in computer prices, making them more affordable.
b) It enables computers to become smaller and more compact in size.
c) It leads to a continuous increase in the cost of electronic components.
d) It facilitates the exponential growth of processing power in computers, making them more powerful and accessible.

2) In which type of machine learning task is the output variable continuous?
a) Classification
b) Regression
c) Clustering
d) Reinforcement learning

3) What is the main objective of clustering algorithms?
a) Finding patterns in unlabeled data
b) Predicting a target variable
c) Classifying data into multiple classes
d) Minimizing the error between predicted and actual values

4) All of the following examples are applications of machine learning, except:
a) Customizing marketing campaigns based on customer demographics and purchase history.
b) Detecting fraudulent activities in financial transactions.
c) Analyzing IoT (Internet of Things) to predict equipment issues before they occur.

5) Which type of ML model is typically described as a "black box"?
a) Linear regression
b) Naive Bayes
c) Artificial neural networks
d) Decision trees

6) Which of the following evaluation metrics is commonly used for classification?
a) Mean Squared Error (MSE)
b) R-squared
c) Accuracy
d) Root Mean Squared Error (RMSE)

7) What is the purpose of cross-validation in machine learning?
a) Evaluating the model's performance on unseen data
b) Splitting the dataset into training and testing sets
c) Tuning hyperparameters to optimize model performance
d) Assessing the feature importance in a model

8) Which statement best explains why neural networks are often referred to as "black boxes"?
a) Modifying a specific weight can have a ripple effect on the network's overall output.
b) Neural networks can have a large number of weights, contributing to their intricate nature.
c) Assigning a direct interpretation to individual weights in a neural network is a challenging task.
d) Each weight in a neural network operates independently, without influencing other neurons.
e) The number of weights in a neural network is directly proportional to its depth.

9) Can neural networks be used to handle structured data?
a) Yes, only structured data
b) No, neural networks are only suitable for unstructured data
c) Yes, both structured and unstructured data
d) No, neural networks cannot handle any type of data

10) Which step of the CRISP-DM process involves assessing data quality, handling missing values, and performing variable transformations?
a) Data understanding
b) Modeling
c) Evaluation
d) Data preparation

For each of the following questions select the correct answer:

1) The EU's Ethics Guidelines for Trustworthy AI were developed to:
   a) Promote the development and adoption of AI in Europe.
   b) Establish strict rules to restrict the use of AI in all domains.
   c) Define principles and values to promote trustworthy and ethical AI.
   d) Facilitate competition among European companies in the field of AI.

2) Among the key principles in the EU's Ethics Guidelines for Trustworthy AI are:
   a) Transparency
   b) Discrimination
   c) Absolute data confidentiality
   d) Unlimited development of AI without restrictions

3) According to the EU's ethical guidelines, AI systems should be designed to:
   a) Respect privacy and protect personal data.
   b) Encourage discrimination and exclusion.
   c) Avoid any interaction with human users.
   d) Use unreliable data to make important decisions.

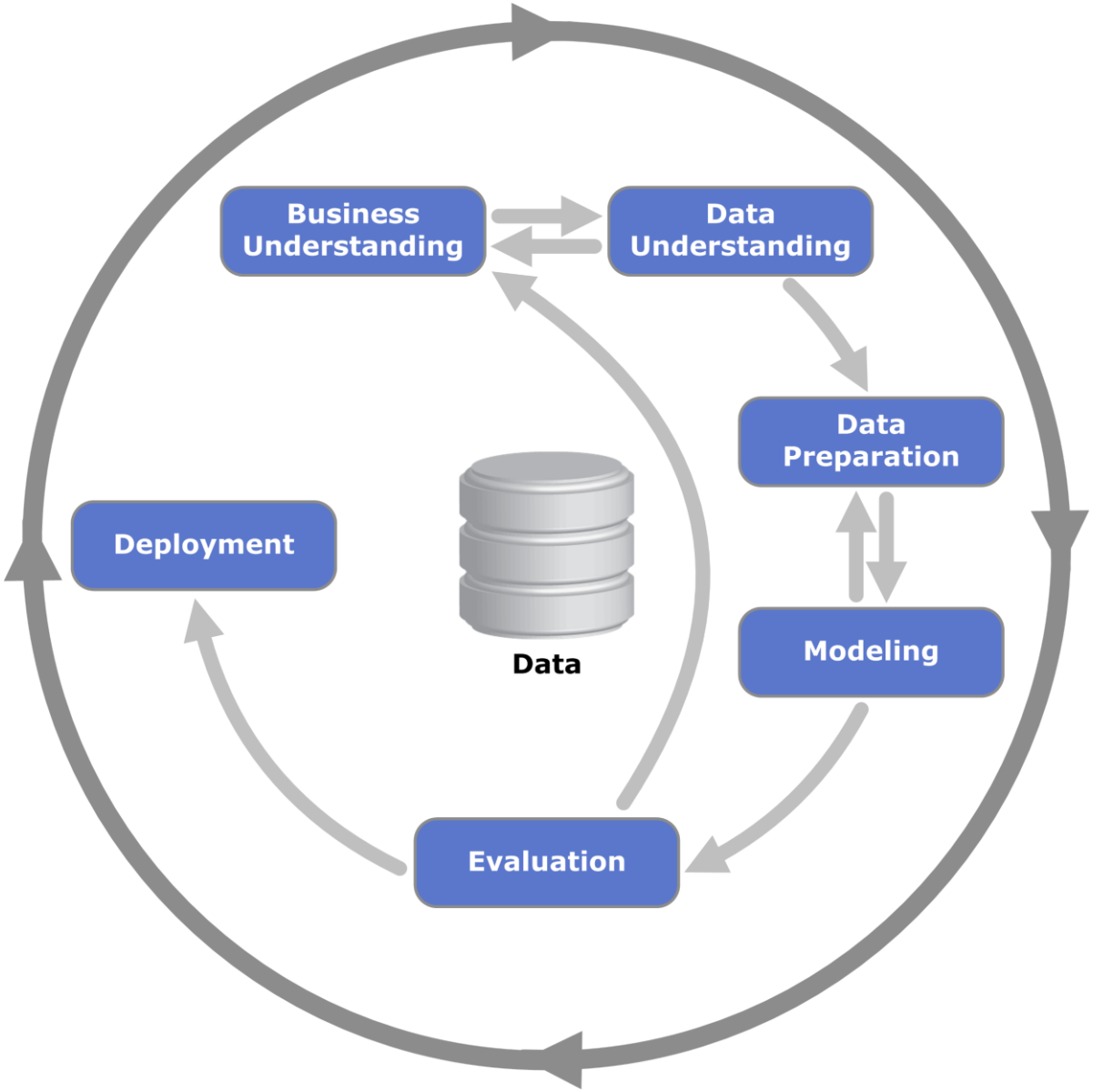
4) The EU's Ethics Guidelines for Trustworthy AI recommend the use of which methods to assess AI-related risks?
   a) No risk assessment is necessary.
   b) Risk assessment based solely on economic criteria.
   c) Multidisciplinary and participatory risk assessment.
   d) Risk assessment based on biases and stereotypes.

5) The EU's Ethics Guidelines for Trustworthy AI emphasize the importance of:
   a) Compliance with existing regulations.
   b) Maximizing profits for AI companies.
   c) Absence of legal liability for AI systems.
   d) Transparency and accountability of AI systems.

Exercice

Rank the above questions according to the stage of the machine learning lifecycle to ensure fairness.

1. Was our historical data generated by a biased process that we treat as real?
2. Can we collect more data or reweight?
3. Is the data distorted?
4. Do our proxies really measure what we think they do?
5. Does our data include enough minority samples?
6. Is the objective function in line with ethics?
7. Do our labels reinforce stereotypes?
8. Are there missing/biased features?
9. Do we need to include fairness constraints in the function?
10. Have we evaluated the model using relevant fairness metrics?
11. Is the algorithm misusable in other contexts?
12. Can we evaluate the model on other datasets beyond the test set?
13. Are we deploying our model on a population that we did not train/test on?
14. Do we need to model minority populations separately?
15. Does the model encourage feedback loops that can produce increasingly unfair outcomes?
16. Is an algorithm an ethical solution to our problem?
17. Do we need to apply debiasing algorithms to preprocess our data?
18. Do our selected fairness metrics capture our customers needs?



Multiple-choice quiz: For each of the following questions, the correct answer(s):

1) What is the primary significance of Moore's Law?
a) It allows for a significant reduction in computer prices, making them more affordable.
b) It enables computers to become smaller and more compact in size.
c) It leads to a continuous increase in the cost of electronic components.
d) It facilitates the exponential growth of processing power in computers, making them more powerful and accessible.
(a) (d)

2) In which type of machine learning task is the output variable continuous?
a) Classification
b) Regression
c) Clustering
d) Reinforcement learning
(b)

3) What is the main objective of clustering algorithms?
a) Finding patterns in unlabeled data
b) Predicting a target variable
c) Classifying data into multiple classes
d) Minimizing the error between predicted and actual values
(a)

4) All of the following examples are applications of machine learning, except:
a) Customizing marketing campaigns based on customer demographics and purchase history.
b) Detecting fraudulent activities in financial transactions.
c) Analyzing IoT (Internet of Things) to predict equipment issues before they occur.
d) Analyzing past revenues to determine the cause of the sales decline.
(c)

5) Which type of ML model is typically described as a "black box"?
a) Linear regression
b) Naive Bayes
c) Artificial neural networks
d) Decision trees
(c)

6) Which of the following evaluation metrics is commonly used for classification?
a) Mean Squared Error (MSE)
b) R-squared
c) Accuracy
d) Root Mean Squared Error (RMSE)
(c)

7) What is the purpose of cross-validation in machine learning?
a) Evaluating the model's performance on unseen data
b) Splitting the dataset into training and testing sets
c) Tuning hyperparameters to optimize model performance
d) Assessing the feature importance in a model
(a) (c)

8) Which statement best explains why neural networks are often referred to as "black boxes"?
a) Modifying a specific weight can have a ripple effect on the network's overall output.
b) Neural networks can have a large number of weights, contributing to their intricate nature.
c) Assigning a direct interpretation to individual weights in a neural network is a challenging task.
d) Each weight in a neural network operates independently, without influencing other neurons.
e) The number of weights in a neural network is directly proportional to its depth.
(b) (c)

9) Can neural networks be used to handle structured data?
a) Yes, only structured data
b) No, neural networks are only suitable for unstructured data
c) Yes, both structured and unstructured data
d) No, neural networks cannot handle any type of data
(c)

10) Which step of the CRISP-DM process involves assessing data quality, handling missing values, and performing variable transformations?
a) Data understanding
b) Modeling
c) Evaluation
d) Data preparation
(d)

For each of the following questions select the correct answer(s):

1) The EU's Ethics Guidelines for Trustworthy AI were developed to:
   a) Promote the development and adoption of AI in Europe.
   b) Establish strict rules to restrict the use of AI in all domains.
   c) Define principles and values to promote trustworthy and ethical AI.
   d) Facilitate competition among European companies in the field of AI.

(c)

2) Among the key principles in the EU's Ethics Guidelines for Trustworthy AI are:
   a) Transparency
   b) Discrimination
   c) Absolute data confidentiality
   d) Unlimited development of AI without restrictions

(a)

3) According to the EU's ethical guidelines, AI systems should be designed to:
   a) Respect privacy and protect personal data.
   b) Encourage discrimination and exclusion.
   c) Avoid any interaction with human users.
   d) Use unreliable data to make important decisions.

(a)

4) The EU's Ethics Guidelines for Trustworthy AI recommend the use of which methods to assess AI-related risks?
   a) No risk assessment is necessary.
   b) Risk assessment based solely on economic criteria.
   c) Multidisciplinary and participatory risk assessment.
   d) Risk assessment based on biases and stereotypes.

(c)

5) The EU's Ethics Guidelines for Trustworthy AI emphasize the importance of:
   a) Compliance with existing regulations.
   b) Maximizing profits for AI companies.
   c) Absence of legal liability for AI systems.
   d) Transparency and accountability of AI systems.

(d)

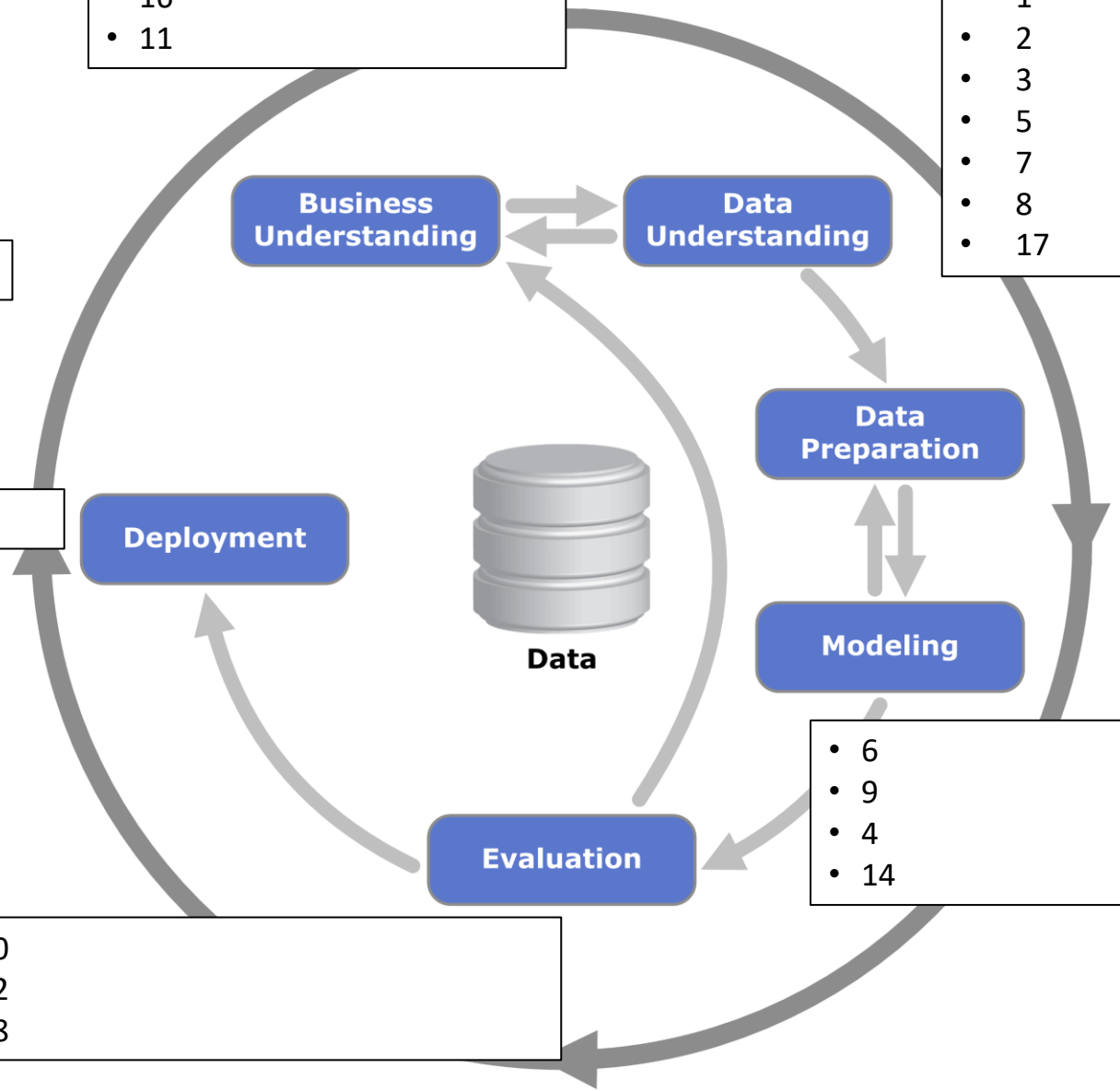
Solution

- 16
- 11

- 1
- 2
- 3
- 5
- 7
- 8
- 17

15

13



- 6
- 9
- 4
- 14

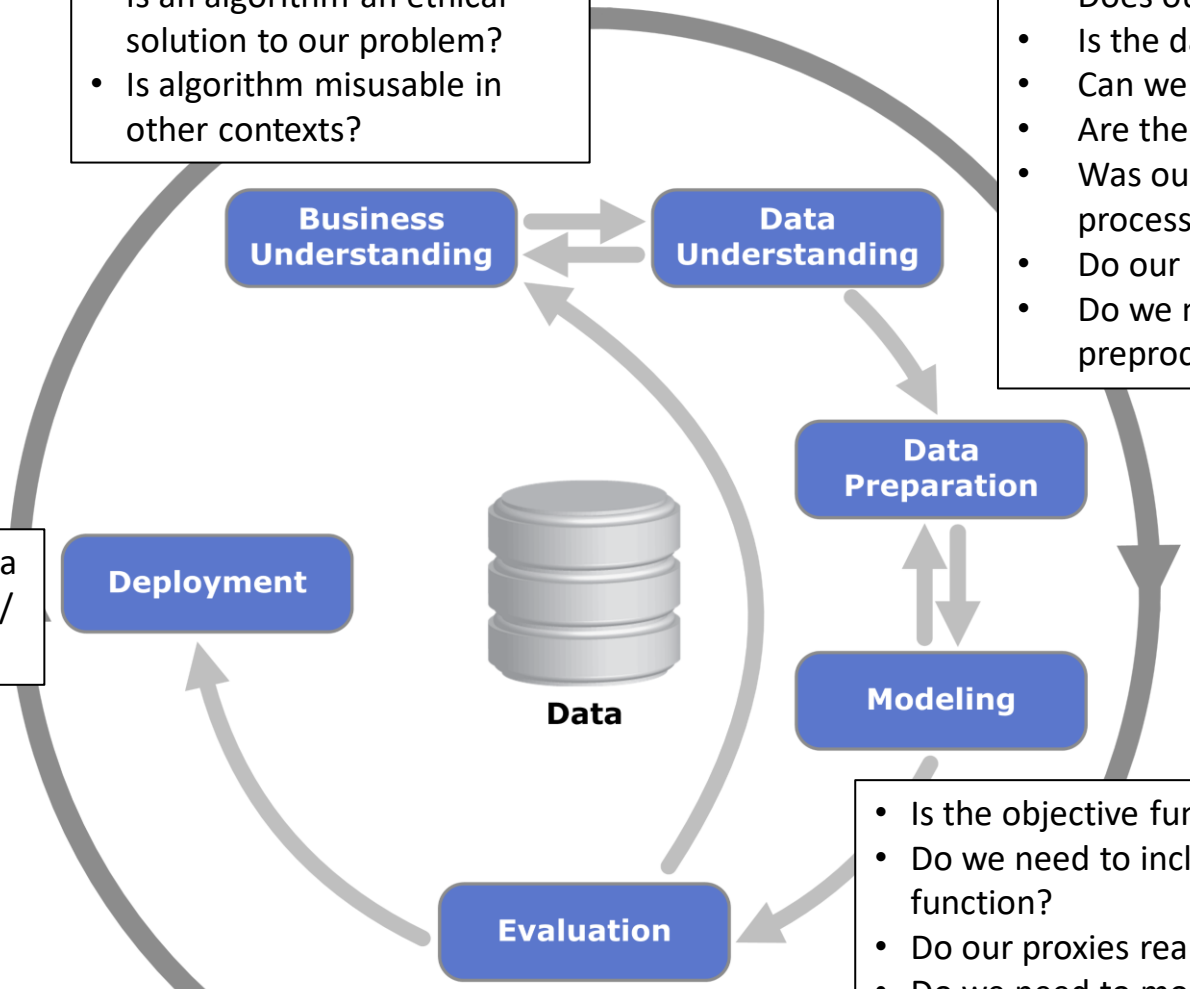
- 10
- 12
- 18

Solution

- Is an algorithm an ethical solution to our problem?
- Is algorithm misusable in other contexts?

- Does our data include enough minority samples?
- Is the data distorted?
- Can we collect more data or reweight?
- Are there missing/biased features?
- Was our historical data generated by a biased processed that we treat as real?
- Do our labels reinforce stereotypes?
- Do we need to apply debiasing algorithms to preprocess our data ?

Does the model encourage feedback loops that can produce increasingly unfair outcomes?



Are we deploying our model on a population that we did not train/ test on?

- Is the objective function in line with ethics?
- Do we need to include fairness constraints in the function?
- Do our proxies really measure what we think they do?
- Do we need to model minority populations separately?

- Have we evaluated the model using relevant fairness metrics?
- Do our selected fairness metrics capture our customers needs?
- Can we evaluate the model on other datasets beyond test set?

Best practices for Bias avoidance/mitigation

1. Practices related to Implementation, monitoring, and awareness
2. Practices related to data preparation and modeling

Best practices for Bias avoidance/mitigation

Practices related to Implementation, monitoring, and awareness

Main Actors :

Human resources managers, Project managers, Training teams, Customer support teams, External auditors

1. Consider team composition for diversity of thought, background, and experiences.
2. Understand the task, stakeholders, and potential for errors and harm.
3. **Post-Deployment**:
   a) Ensure optimization and guardrail metrics consistent with responsible practices and avoid harms.
   b) Continual monitoring, including customer feedback.
   c) Have a plan to identify and respond to failures and harms as they occur.
4. **Conduct external audits or third-party evaluations**: Seek external assessments or audits of the AI system to gain independent insights into potential biases and ensure compliance with ethical and legal standards.
5. **Foster ongoing education and awareness**: Promote awareness and understanding of AI biases among team members, stakeholders, and users. Encourage ongoing education and training on responsible AI practices and the potential impacts of biases.

Best practices for Bias avoidance/mitigation

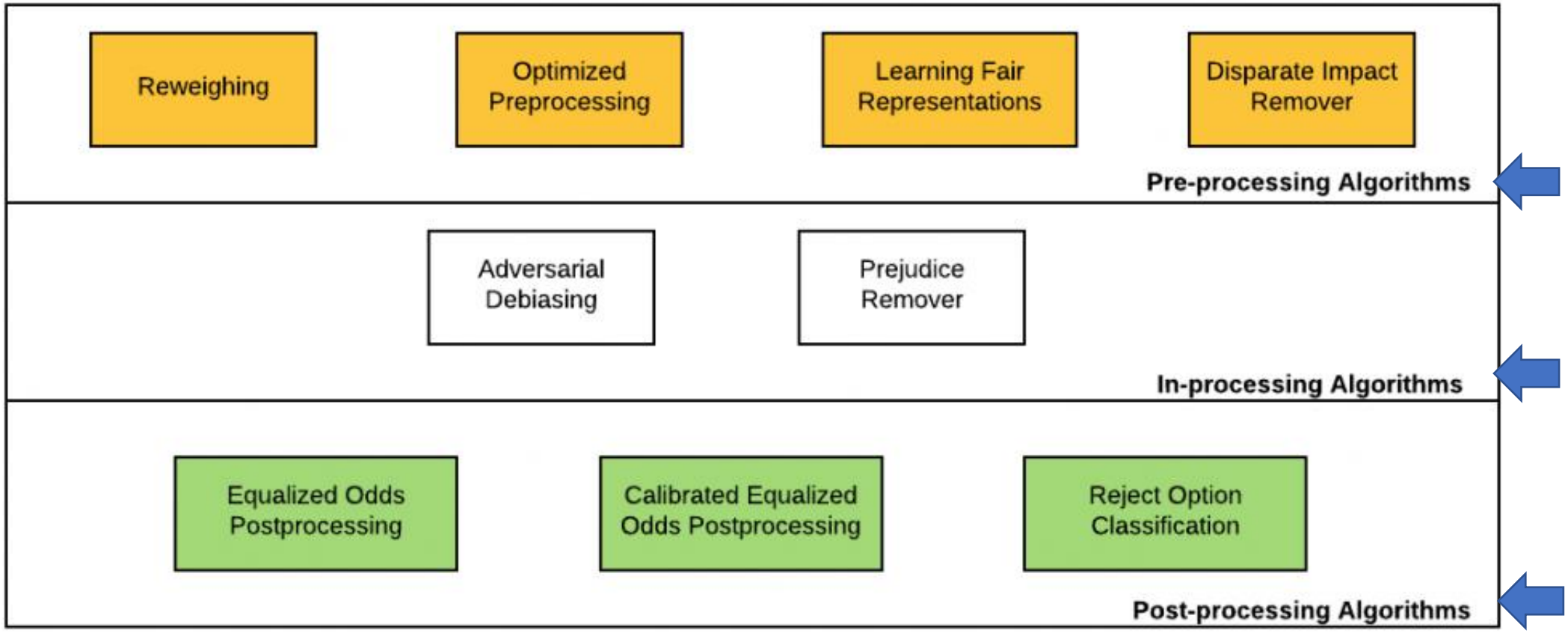
Practices related to data preparation and modeling

Main Actors : Data scientists, Machine learning engineers, Domain experts

1. **Check data sets:**
    a) Consider data provenance.
    b) Verify the data using qualitative, experimental, survey, and other relevant methods.
2. **Apply Bias Mitigation Strategies for ML models**
    a) Pre-processing algorithms.
    b) In-processing algorithms.
    c) Post-processing algorithms.

Bias Mitigation Strategies for ML models

Some practices related to data preparation and modeling :



https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies
https://towardsdatascience.com/approaches-for-addressing-unfairness-in-machine-learning-a31f9807cf31

Bias Mitigation Strategies for ML models

Pre-Processing Algorithms

- **Reweighing**: Reweighing is a data preprocessing technique that recommends generating weights for the training examples in each (group, label) combination differently to ensure fairness before classification. The idea is to apply appropriate weights to different tuples in the training dataset to make the training dataset discrimination free with respect to the sensitive attributes : Weighting enables positive discrimination.
- **Optimized preprocessing**: The idea is to learn a probabilistic transformation that modifies the features and labels of data while respecting constraints and objectives related to group fairness, individual distortion, and data fidelity.
- **Learning fair representations**: The idea is to find a latent representation that encodes the data well while hiding information about protected attributes. For example In the context of recruitment, the model is trained using a dataset that includes candidate information, such as gender or ethnicity, but once it learns to generate the latent representation, it may focus on relevant factors like skills, experience, and qualifications rather than relying on gender or ethnicity to make hiring decisions.
- **Disparate impact remover**: Feature values are appropriately modified to increase group fairness while preserving rank-ordering within groups (statistical approach). For example, let's say you have a university admission prediction model that exhibits unfair treatment between male and female applicants. The "Disparate Impact Remover" would analyze the features used by the model (such as test scores, grades, etc.) and adjust the values of those features to reduce the treatment disparities. The goal is to achieve increased fairness between the groups while preserving the rank ordering of the candidates within each group.

Bias Mitigation Strategies for ML models

In-Processing Algorithms

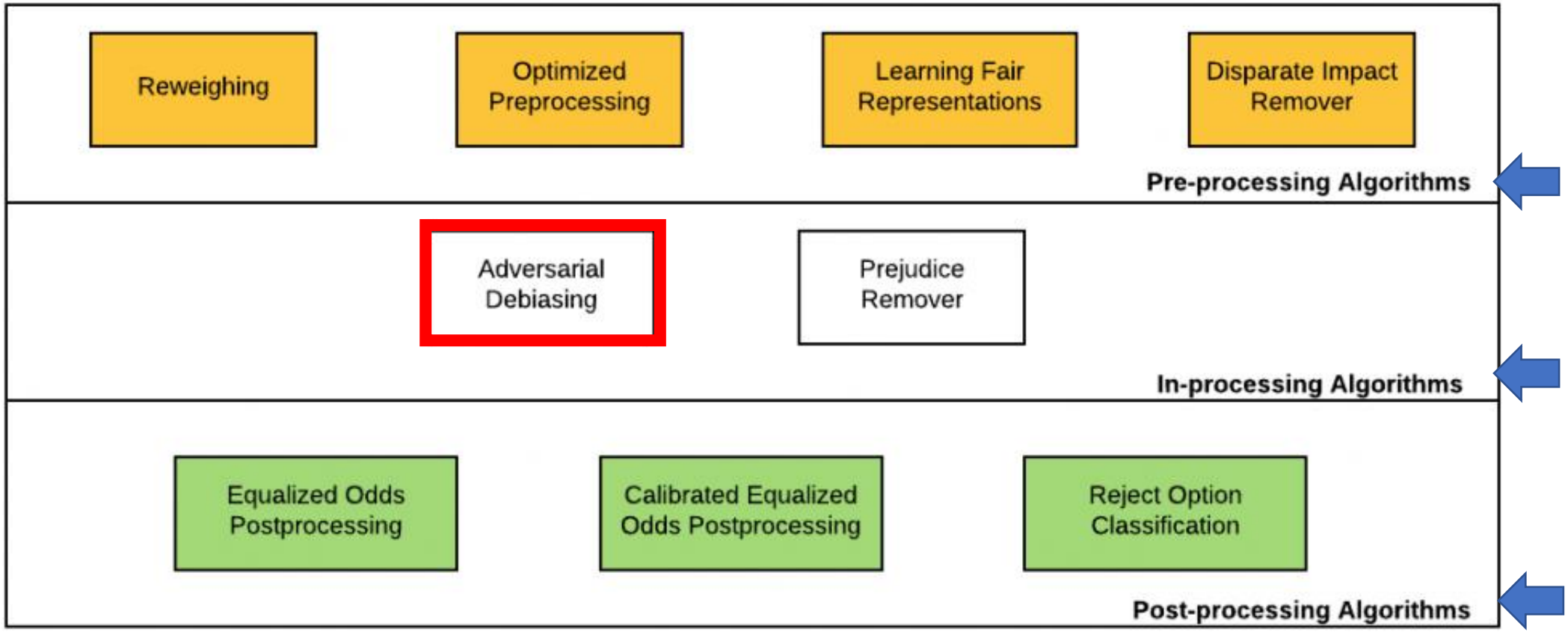
- **Adversarial Debiasing**: A classifier model is learned to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions [**to be developed**]
- **Prejudice remover**: The idea is to add a discrimination-aware regularization term to the learning objective.

Post-Processing Algorithms

- **Equalized odds postprocessing**: The algorithm solves a linear program to find probabilities with which to change output labels to optimize equalized odds.
- **Calibrated equalized odds postprocessing**: The algorithm optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.
- **Reject option classification**: The idea is to give favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary

Bias Mitigation Strategies for ML models

Some of the bias mitigation strategies that can be applied in ML Model Development lifecycle (MDLC) to achieve discrimination-aware Machine Learning models:



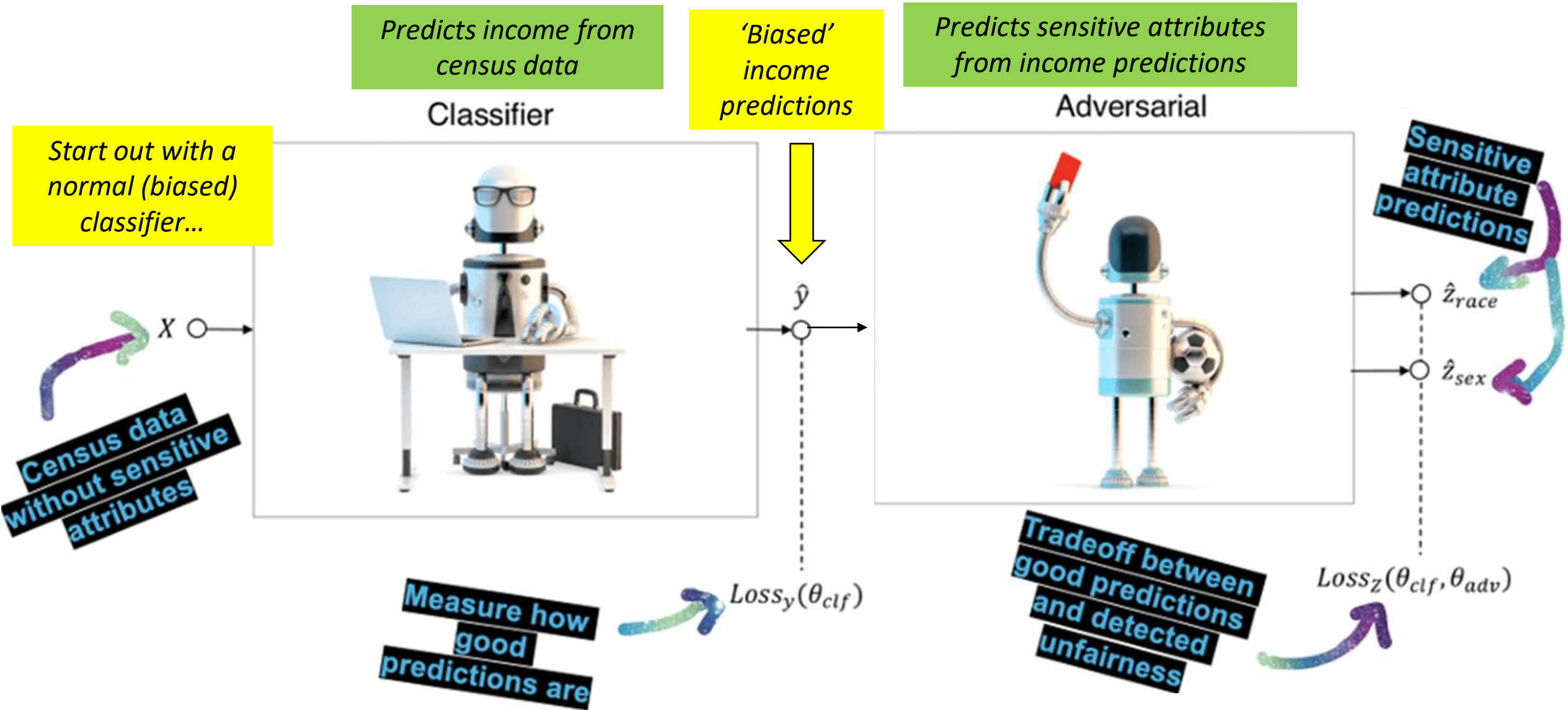
https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies

Exemple of how make fair machine learning models

Adversarial Debiasing

Training for fairness : Adversarial training procedure

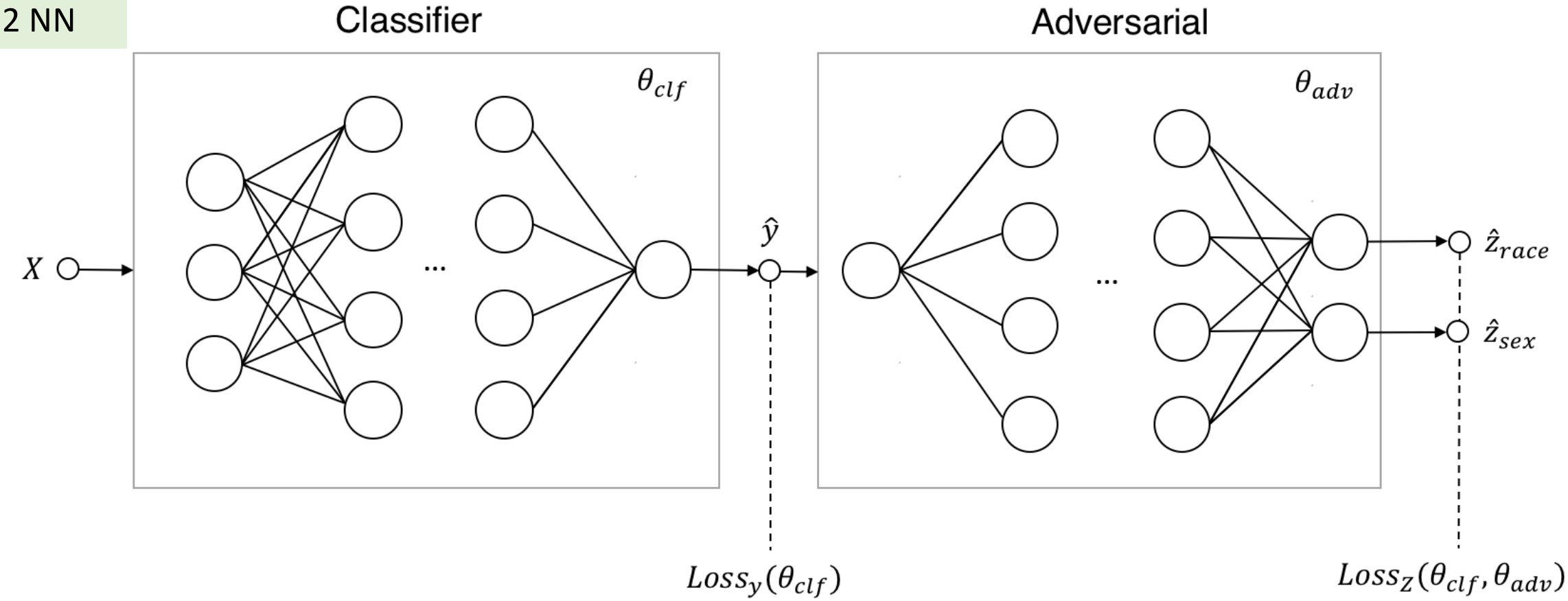
- **Principle**: Enforce fairness by adding an adversarial classifier as a fairness referee to a normal (biased) classifier
- This referee tries to reconstruct bias from the predictions and penalizes the classifier if it can find any unfairness.



Training for fairness : Adversarial training procedure

- Principle: Enforce fairness by adding an adversarial classifier as a fairness referee to a normal (biased) classifier
- This referee tries to reconstruct bias from the predictions and penalizes the classifier if it can find any unfairness.

use **PyTorch** with 2 NN



Example: "Census Income"

fairness-in-torch.ipynb

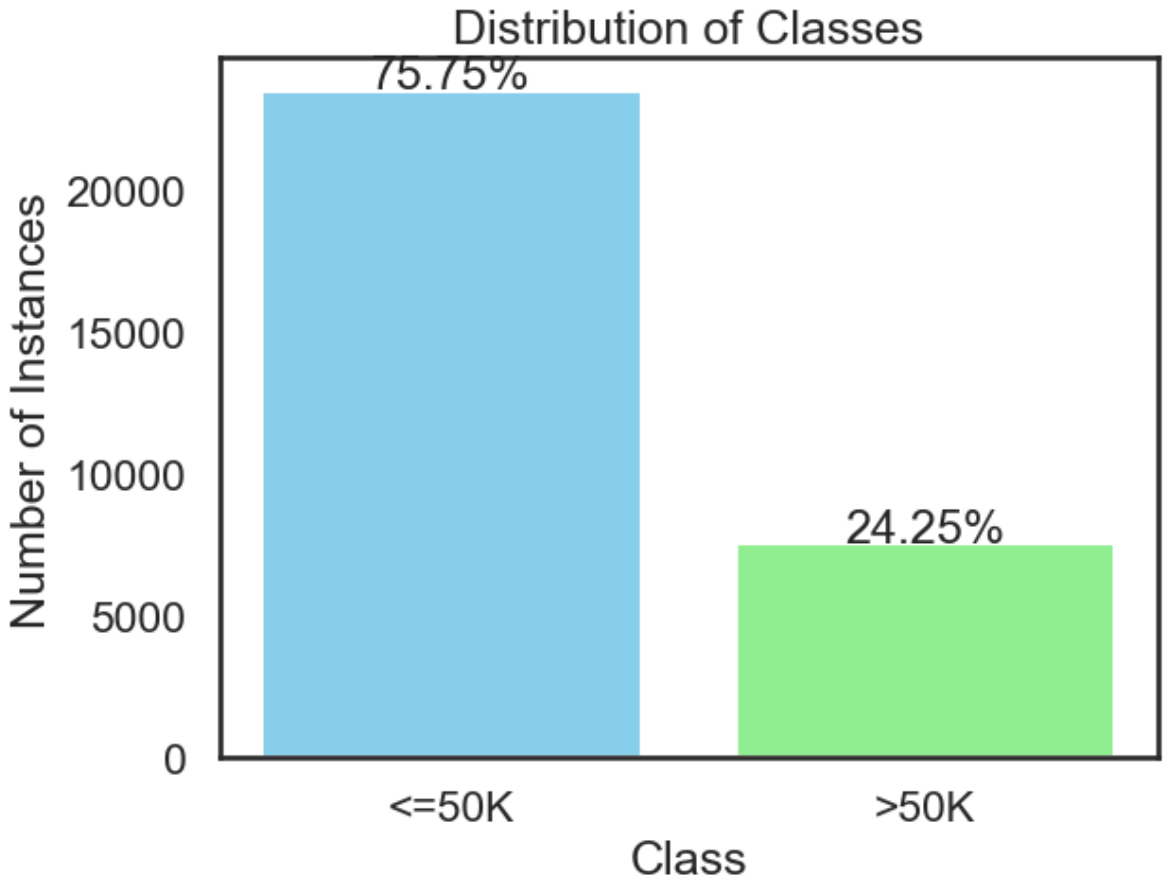
- Predict **income level** using **Adult dataset** (https://archive.ics.uci.edu/ml/datasets/Adult) that involves predicting personal income levels as above or below $50,000 per year based on personal details

  # Instances 48842 **# Attributes** 14
- The used approach is based on the 2017 NIPS paper "Learning to Pivot with Adversarial Networks" [Louppe et al., 2017

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	target
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-abs	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Isl	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Isl	Male	0	0	40	?	>50K

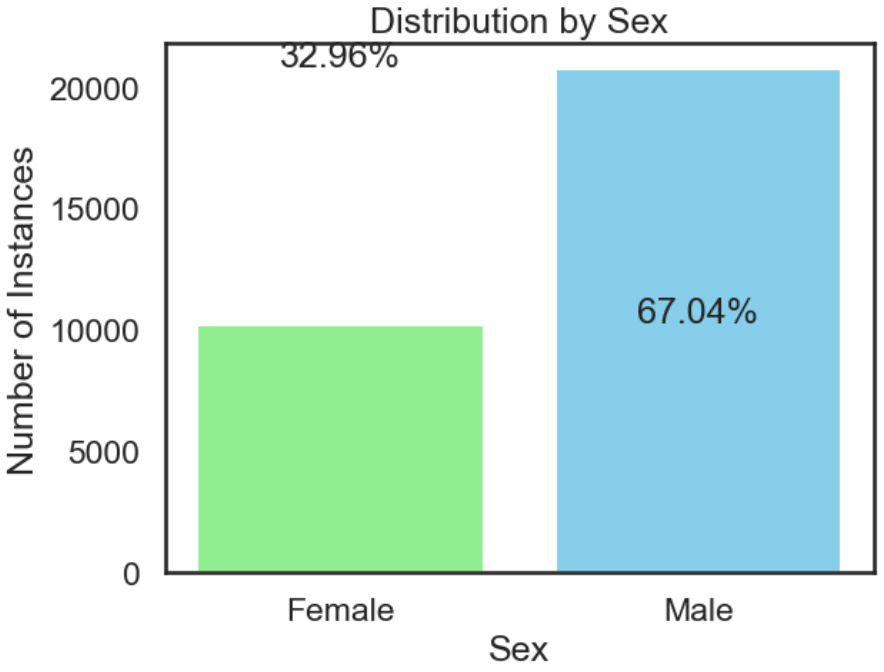
1. **The set of features** contains the input attributes that the model uses for making the predictions, with attributes like age, education level and occupation.
2. **The target** contains the binary class that the model needs to predict (above or below $50K).

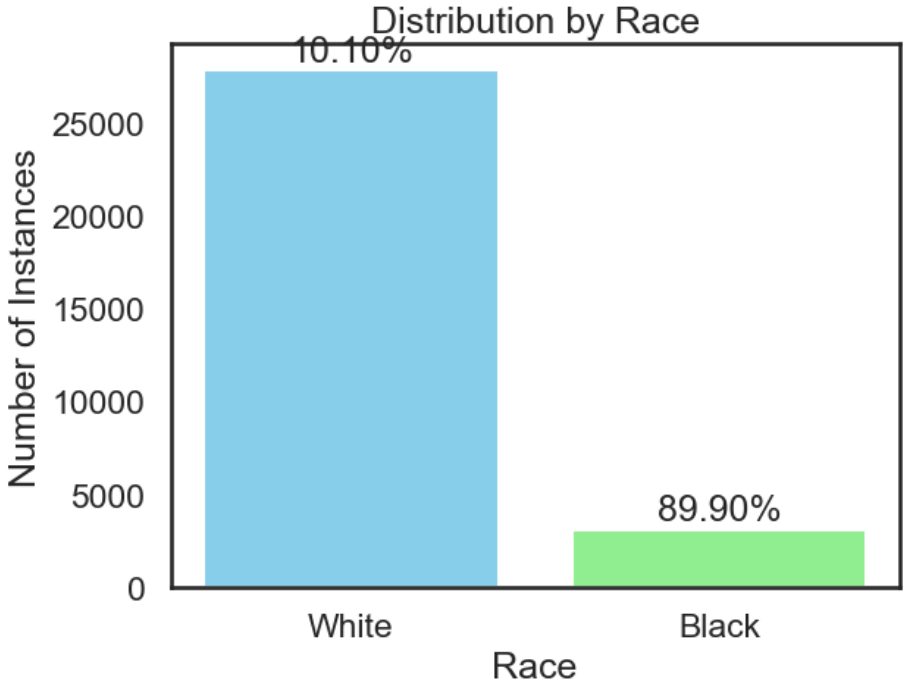
Take a look on the data…



- Initial observation: The dataset exhibits class **imbalance**, meaning that the distribution of class labels is skewed, with one class (low-income class) having significantly more instances than the other.
- However, addressing this class imbalance is not our current focus.

Take a look on the data…





- We consider two sensitive attributes, **Sex** and **Race**, have imbalanced distributions that can lead to unfair decisions
- How to measure fairness and how make fair machine learning models ?

Example: "Census Income" dataset

We dispatch the data into 3 subsets:

1. **The set of features** X contains the input attributes that the model uses for making the predictions, with attributes like age, education level and occupation.
2. **The targets** y contain the binary class labels that the model needs to predict. These labels are $y \in \{income > 50K, income \leq 50K\}$
3. **The set of sensitive attributes** Z contains the attributes for which we want the prediction to be fair. These are $zrace \in \{black, white\}$ and $zsex \in \{male, female\}$

y

X Z X

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	target
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-famil	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-famil	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-abs	Other-service	Not-in-famil	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-not	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-famil	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Isl	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-famil	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Isl	Male	0	0	40	?	>50K

Example: "Census Income" dataset

…/fairness-in-torch.ipynb

```
 load ICU data set
X, y, Z = load_ICU_data('data/adult.data')
```

…/fairness/helpers.py

```
Z = (input_data.loc[:, sensitive_attribs]
   .assign(race=lambda df: (df['race'] == 'White').astype(int),
    sex=lambda df: (df['sex'] == 'Male').astype(int)))

   # targets; 1 when someone makes over 50k , otherwise 0
   y = (input_data['target'] == '>50K').astype(int)

   # features; 'target' and sentive attribute columns are dropped
   X = (input_data
  .drop(columns=['target', 'race', 'sex', 'fnlwgt'])
  .fillna('Unknown')
  .pipe(pd.get_dummies, drop_first=True))
```

One-hot encoding : converts categorical variables into distinct binary variables, indicating the presence or absence of each category – thus we pass from 14 to 93 attributes

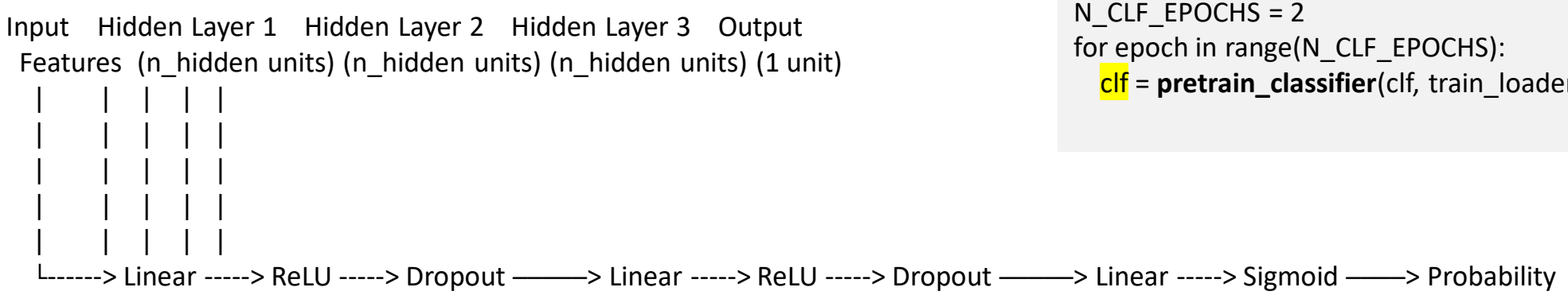
```
features X: 30940 samples, 93 attributes
targets y: 30940 samples
sensitives Z: 30940 samples, 2 attributes
```

- It is important to note that datasets are non-overlapping, so the sensitive attributes **race** and **sex** are not part of the features used for training the model.

Classifier (1)

1. We train a basic income level predictor using **PyTorch**.
2. The network consists of **three** sequential hidden layers with ReLu activation and dropout. The sigmoid layer turns these activations into a probability for the income class (i.e. probability of income being greater than 50K).

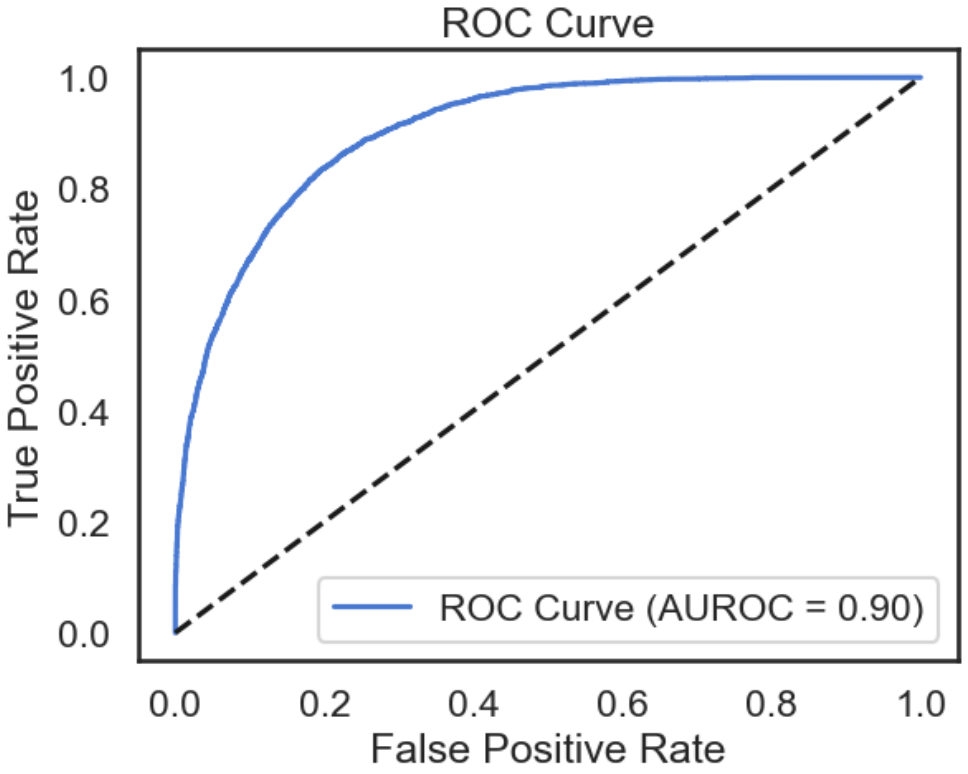
```
class Classifier(nn.Module):

  def __init__(self, n_features, n_hidden=32, p_dropout=0.2):
 super(Classifier, self).__init__()
 self.network = nn.Sequential(
    nn.Linear(n_features, n_hidden),
    nn.ReLU(),
    nn.Dropout(p_dropout),
    nn.Linear(n_hidden, n_hidden),
    nn.ReLU(),
    nn.Dropout(p_dropout),
    nn.Linear(n_hidden, n_hidden),
    nn.ReLU(),
    nn.Dropout(p_dropout),
    nn.Linear(n_hidden, 1),
 )

  def forward(self, x):
 return torch.sigmoid(self.network(x))
…

N_CLF_EPOCHS = 2
for epoch in range(N_CLF_EPOCHS):
  clf = pretrain_classifier(clf, train_loader, clf_optimizer, clf_criterion)
```



Classifier (2)

```
#Evaluate the accuracy of the classifier on the test data
test_accuracy = evaluate_classifier(clf, test_data)
print("Accuracy on test data:", test_accuracy)
```

```
Accuracy on test data: 0.85%
```



- The ROC (Receiver Operating Characteristic) curve is a graphical representation used to evaluate the performance of a binary classification model.
- It displays the relationship between the true positive rate and the false positive rate at different classification thresholds, which are probability thresholds that determine from which point an example is considered positive or negative.
- The closer the ROC curve is to the upper-left corner of the graph, the better the model is in terms of classification performance.
- Area Under the (AUROC): measures the ability of the model to distinguish between positive and negative classes by calculating the area under the (ROC) curve.

With a AUCROC =0.9 and a prediction accuracy of 85% we can say that the basic classifier performs pretty well!

However, if it is also fair in its predictions, that remains to be seen…

How to measure fairness ?

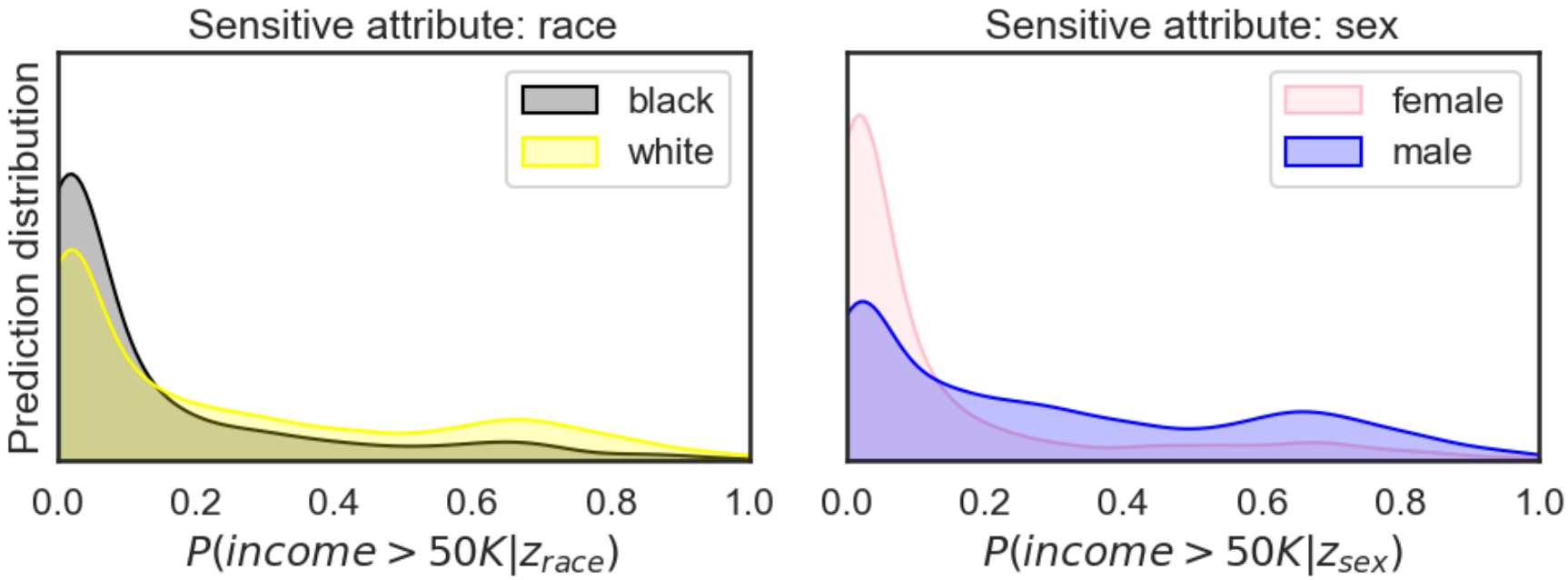
- Qualitative model
- Quantitative model

Qualitative model fairness

- We start the investigation into the fairness of our classifier by analysing the predictions it made on the test set.
- The plots in the figure below show the distributions of the predicted $P(income>50K)$ given the sensitive attributes.

```
with torch.no_grad():
    pre_clf_test = clf(test_data.tensors[0])
```

- **pre_clf_test** stores the raw predictions of the clf model on the test data (transformed into y_pre_clf in the code)



Individuals who are classified as black and/or female have a significantly higher likelihood of being predicted to have an income below 50K compared to those who are white and/or male.

- **The predictions are biased when considered in the context of race and sex.**
- The model tends to favor white males when it comes to assigning high-income levels

Quantitative model fairness

Here is a sample list of fairness criteria

Name	Reference
Statistical parity	Dwork et al. (2011)
Group fairness	
Demographic parity	
Conditional statistical parity	Corbett-Davies et al. (2017)
Darlington criterion (4)	Darlington (1971)
Equal opportunity	Hardt, Price, Srebro (2016)
Equalized odds	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Berk et al. (2017)
Avoiding disparate mistreatment	Zafar et al. (2017)
Balance for the negative class	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Chouldechova (2016)
Equalized correlations	Woodworth (2017)
Darlington criterion (3)	Darlington (1971)
Cleary model	Cleary (1966)
Conditional use accuracy	Berk et al. (2017)
Predictive parity	Chouldechova (2016)
Calibration within groups	Chouldechova (2016)
Darlington criterion (1), (2)	Darlington (1971)

https://fairmlbook.org/pdf/fairmlbook.pdf

Demographic Parity

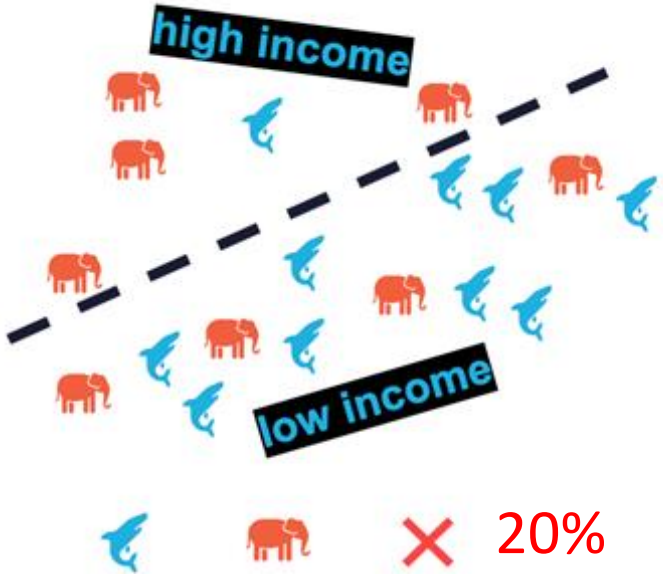
- In order to get a 'quantitative' measure of how fair our classifier is, we use the **Demographic Parity** (p%-rule), also called **Statistical Parity** or **Equalizing acceptance rate**, it is one of the most well-known criteria for fairness.
- A classifier that makes binary class predictions ($\hat{y} \in \{0,1\}$) given a binary sensitive attribute ($z \in \{0,1\}$), such that "z=1" represents the sensitive attribute being true (disadvantaged group) and "z=0" represents the sensitive attribute being false (advantaged group), will have the same acceptance rate if the probability of a positive outcome ($\hat{y} = 1$) is equal for both groups:

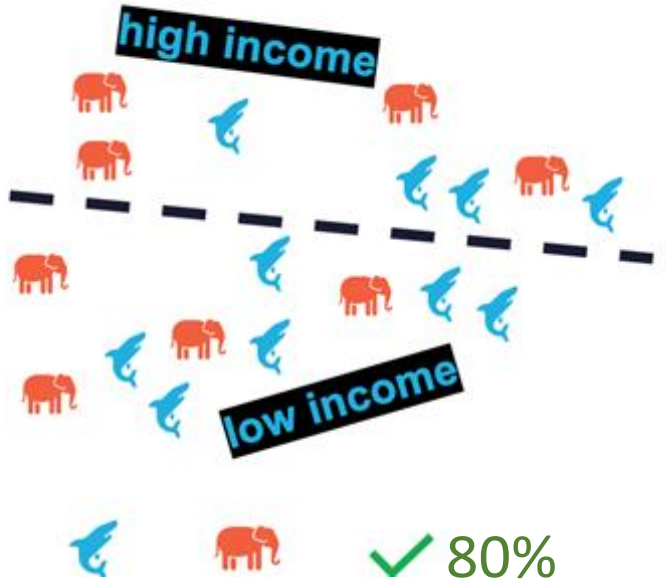
$$p(\hat{y} = 1|z = 1) = p(\hat{y} = 1|z = 0)$$

- Sometimes, we allow for some flexibility. The 80%-rule states that the chance of acceptance for the disadvantaged group should be at least within 80% of the other group (ε=0.2):

$$\frac{p(\hat{y} = 1|z = 1)}{p(\hat{y} = 1|z = 0)} \geq 0.8$$

- When a classifier is completely fair, it will satisfy a 100%-rule. In contrast, when it is completely unfair, it satisfies a 0%-rule.





- Acceptance rate for dolphins: (number of dolphins with a positive prediction) / (total number of dolphins) = 1 / 11 = 0.0909
- Acceptance rate for elephants: (number of elephants with a positive prediction) / (total number of elephants) = 4 / 9 = 0.4444
- Ratio of acceptance rates: rate) = 0.0909 / 0.4444 = 0.2045

- Acceptance rate for dolphins: (number of dolphins with a positive prediction) / (total number of dolphins) = 4 / 11= 0.3636
- Acceptance rate for elephants: (number of elephants with a positive prediction) / (total number of elephants) = 4/ 9 = 0.4444
- Ratio of acceptance rates: = 0.3636 / 0.4444 = 0.8182

```
…/fairness/helpers.py
def p_rule(y_pred, z_values, threshold=0.5):
    y_z_1 = y_pred[z_values == 1] > threshold if threshold else y_pred[z_values == 1]
    y_z_0 = y_pred[z_values == 0] > threshold if threshold else y_pred[z_values == 0]
    odds = y_z_1.mean() / y_z_0.mean()
    return np.min([odds, 1/odds]) * 100
```

- **threshold** : used to transform probabilities into binary values (0/1)
- **odds** calculates the ratio between :
  average of positive predictions in the `z_values == 1` group
  average of positive predictions in the `z_values == 0` group.

- We select the smaller value between 'odds' and '1/odds' to ensure that the ratio of acceptance rates is always less than or equal to 1.

```
…/fairness/helpers.py
p_rules = {'race': p_rule(y_pred, Z_test['race']),
           'sex' : p_rule(y_pred, Z_test['sex']),}
```

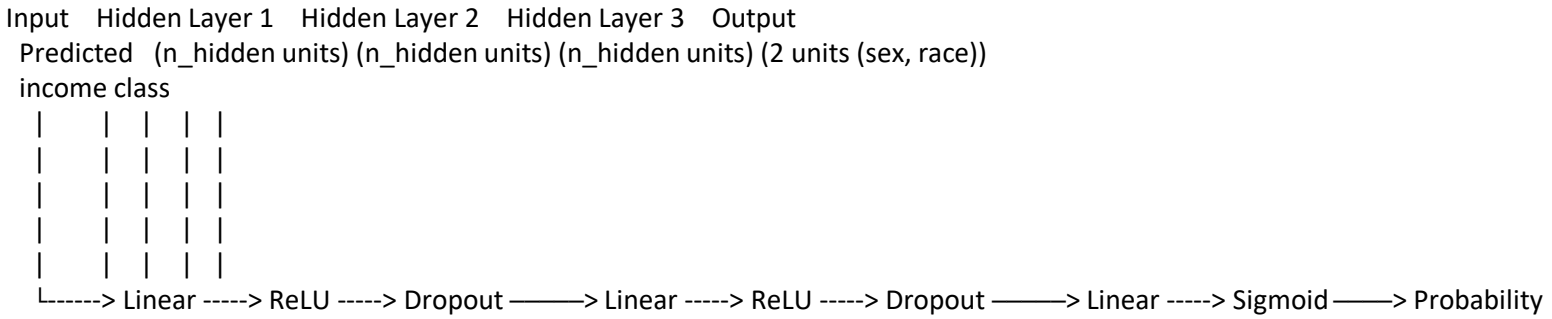
Satisfied p%-rules:
- race: 42%-rule
- sex: 39%-rule

- For both sensitive attributes the classifier satisfies a p%-rule that is significantly lower than 80%.
- This supports our earlier conclusion that the trained classifier is unfair in making its predictions.

Adversary

```
class Adversary(nn.Module):

   def __init__(self, n_sensitive, n_hidden=32):
 super(Adversary, self).__init__()
 self.network = nn.Sequential(
     nn.Linear(1, n_hidden),
     nn.ReLU(),
     nn.Linear(n_hidden, n_hidden),
     nn.ReLU(),
     nn.Linear(n_hidden, n_hidden),
     nn.ReLU(),
     nn.Linear(n_hidden, n_sensitive),
 )

   def forward(self, x):
 return torch.sigmoid(self.network(x))
```

- The adversary (adv) has the same structure than the classifier (clf)
- The input comes from a single class (the predicted income class) and the output consists of two sensitive classes (sex and race).




```
…
N_ADV_EPOCHS = 5

for epoch in range(N_ADV_EPOCHS):
   adv= pretrain_adversary(adv, clf, train_loader, adv_optimizer,  adv_criterion)
…
with torch.no_grad():
   pre_adv_test = adv(pre_clf_test)
```

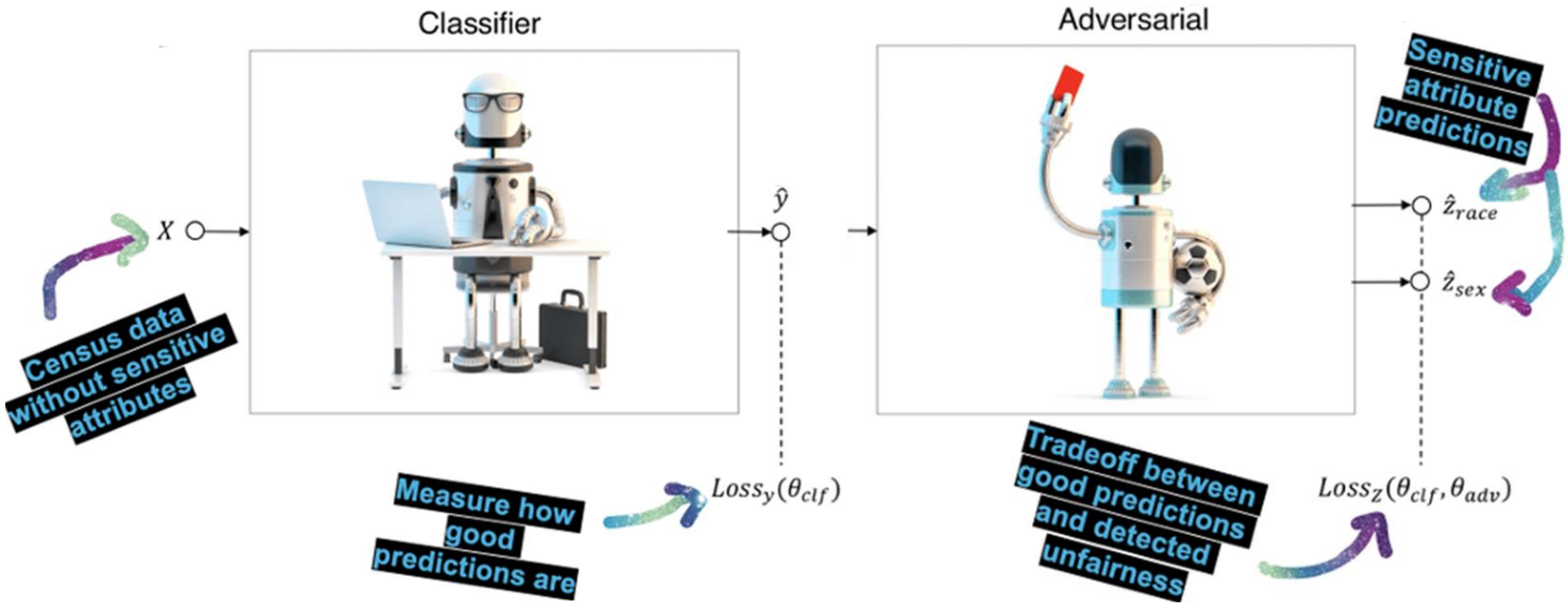
Adversary performance:
- ROC AUC: 0.66

- **pre_adv_test** stores the raw predictions of the adv model on the test data (transformed into y_pre_adv in the code)

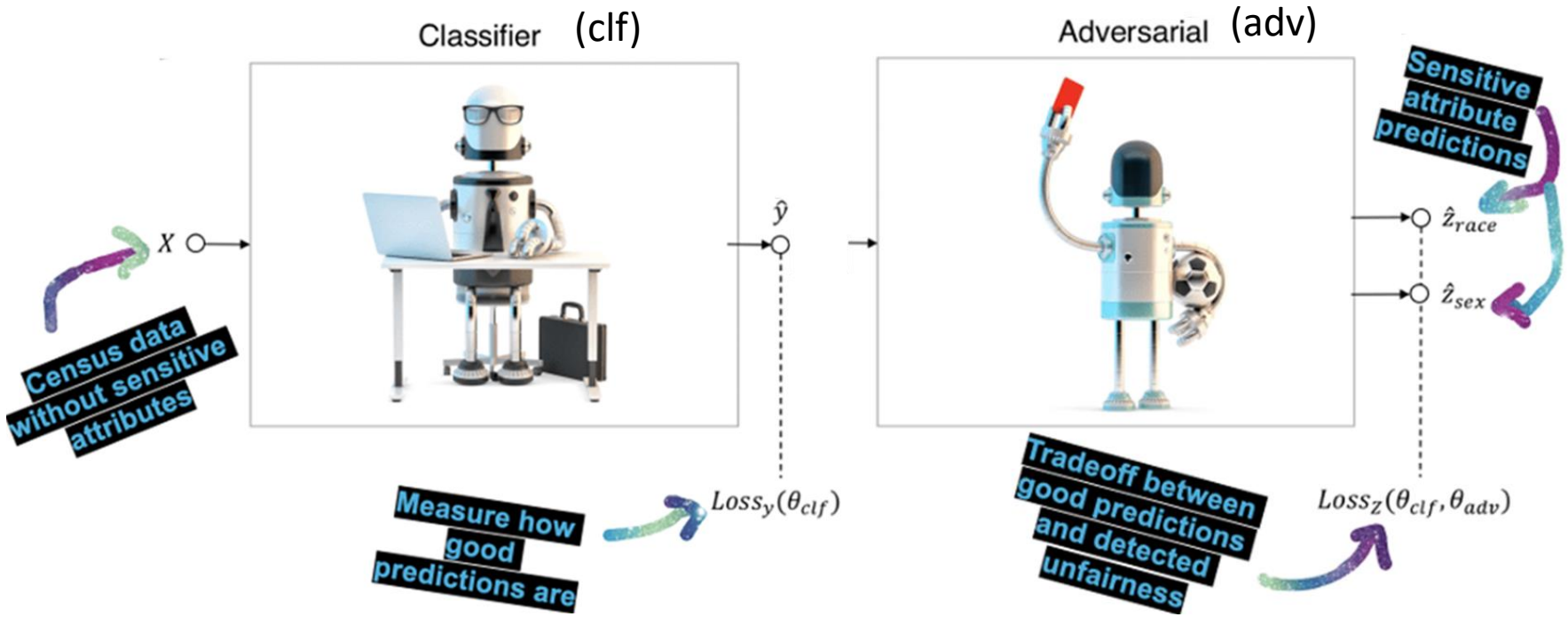
Training for fairness

Now that we have an unfair classifier and an adversary that is able to pick up on unfairness, we can engage them in the **zero-sum game** to make the classifier fair.

- The zero-sum game consists of a competition between the classifier and the adversary, where both parties are in conflict.
  - The classifier aims to make accurate predictions on the data,
  - while the adversary seeks to detect unfair decisions made by the classifier regarding sensitive attributes.



Training for fairness

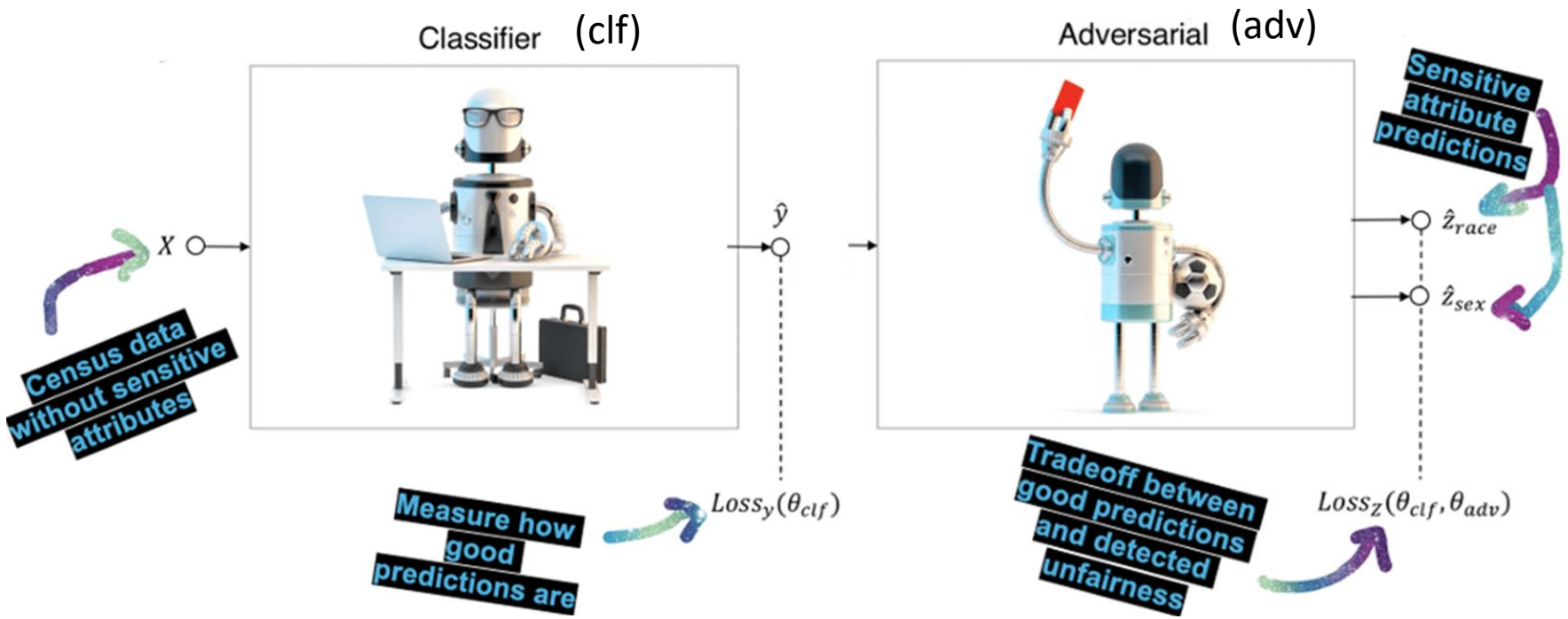


- **The classifier (clf) aims** to minimize its prediction losses ($Loss_y$) while also minimizing the ability of the adversary to detect unfair decisions ($Loss_Z$) by adjusting its parameters (θ_{clf}).
- The objective function for the classifier is:

$$\min_{\theta_{clf}}[Loss_y(\theta_{clf}) - \lambda\, Loss_Z(\theta_{clf}, \theta_{adv})]$$

λ represents a scalar parameter that allows adjusting the relative importance of the adversary's loss compared to the classifier's loss in the optimization problem.

Training for fairness



- **The adversary (adv)** aims to maximize its ability to detect instances where the classifier's decisions are unfair with respect to sensitive attributes.
- The loss function captures the discrepancy between the predictions of the classifier (based on θ_{clf}) and the adversary (based on θ_{adv}) regarding unfair decisions.
- The objective function of the adversary is to minimize the loss function $Loss_Z(\theta_{clf},\theta_{adv})$, which measures its ability to detect unfair decisions made by the classifier:

$$\min_{\theta_{adv}}[Loss_Z(\theta_{clf},\theta_{adv})]$$

Training for fairness

We can summarize this procedure in the following 3 steps:

1. Pre-train the classifier on the full data set. (DONE)
2. Pre-train the adversarial on the predictions of the pre-trained classifier. (DONE)

 # The actual adversarial training starts only after the first two pre-training steps

3. During T iterations simultaneously train the adversarial and classifier networks:
   - first train the adversarial for a single epoch while keeping the classifier fixed
   - then train the classifier on the full dataset for several epochs while keeping the adversarial fixed.

   (Note that originally training was done on a single random minibatch, because this approach greatly speeds up the training procedure.)

```
def train(clf, adv, data_loader, clf_criterion, adv_criterion, clf_optimizer, adv_optimizer, lambdas):

    # Train adversary
    for x, y, z in data_loader:
        p_y = clf(x)
        adv.zero_grad()
        p_z = adv(p_y)
        loss_adv = (adv_criterion(p_z, z) * lambdas).mean()
        loss_adv.backward()
        adv_optimizer.step()

    # Train classifier on single batch (pass in for)
    for x, y, z in data_loader:

    p_y = clf(x)
    p_z = adv(p_y)
    clf.zero_grad()
    clf_loss = clf_criterion(p_y, y) - (adv_criterion(adv(p_y), z) * lambdas).mean()
    clf_loss.backward()
    clf_optimizer.step()

    return clf, adv
```

Training for fairness

```
  # Train adversary
 for x, y, z in data_loader:
     p_y = clf(x)
     adv.zero_grad()
     p_z = adv(p_y)
     loss_adv = (adv_criterion(p_z, z) * lambdas).mean()
     loss_adv.backward()
     adv_optimizer.step()
```

lambdas = torch.Tensor([130, 30])

- 130 : optimization process will give higher importance to minimizing the adversary's loss.
- 30: optimization process will assign less importance to minimizing the classifier's loss.

1. **p_y = clf(x)**: The classifier model (**clf**) predicts the income (**p_y**) based on the input **x**.
2. **adv.zero_grad()**: The gradients of the adversary model (**adv**) are reset to zero.
3. **p_z = adv(p_y)**: The adversary model makes predictions (**p_z**) based on the income predictions (**p_y**) from the classifier model.
4. **loss_adv =** The loss between the adversary's predictions (**p_z**) and the true labels (**z**) is calculated using the adversary criterion (**adv_criterion**). The loss is then multiplied by a scalar value **lambdas** and averaged across the batch. (objective function of adv)
5. **loss_adv.backward()**: The gradients of the loss w.r.t. the adversary model's parameters are computed (minimize the objective function)
6. **adv_optimizer.step()**: update the parameters of the adversary model (**θ_adv**).

```
def train(clf, adv, data_loader, clf_criterion, adv_criterion, clf_optimizer, adv_optimizer, lambdas):

    # Train adversary
  for x, y, z in data_loader:
        p_y = clf(x)
        adv.zero_grad()
        p_z = adv(p_y)
        loss_adv = (adv_criterion(p_z, z) * lambdas).mean()
        loss_adv.backward()
        adv_optimizer.step()

    # Train classifier on single batch
    for x, y, z in data_loader:

    p_y = clf(x)
    p_z = adv(p_y)
    clf.zero_grad()
    clf_loss = clf_criterion(p_y, y) - (adv_criterion(adv(p_y), z) * lambdas).mean()
    clf_loss.backward()
    clf_optimizer.step()

    return clf, adv
```

Training for fairness

```
# Train classifier on single batch
for x, y, z in data_loader:

p_y = clf(x)
p_z = adv(p_y)
clf.zero_grad()
clf_loss = clf_criterion(p_y, y) - (adv_criterion(adv(p_y), z) * lambdas).mean()
clf_loss.backward()  #minimize the objectove function
clf_optimizer.step()  #update (θ_clf)
```

1. **p_y = clf(x)**: The classifier model (**clf**) predicts the income (**p_y**) based on the input **x**.
2. **p_z = adv(p_y)**: The adversary model (**adv**) makes predictions (**p_z**) based on the income predictions (**p_y**) from the classifier model.
3. **clf.zero_grad()**: The gradients of the classifier model (**clf**) are reset to zero.

The objective function for training the classifier.

clf_loss = clf_criterion(p_y, y) - (adv_criterion(adv(p_y), z) * lambdas).mean()

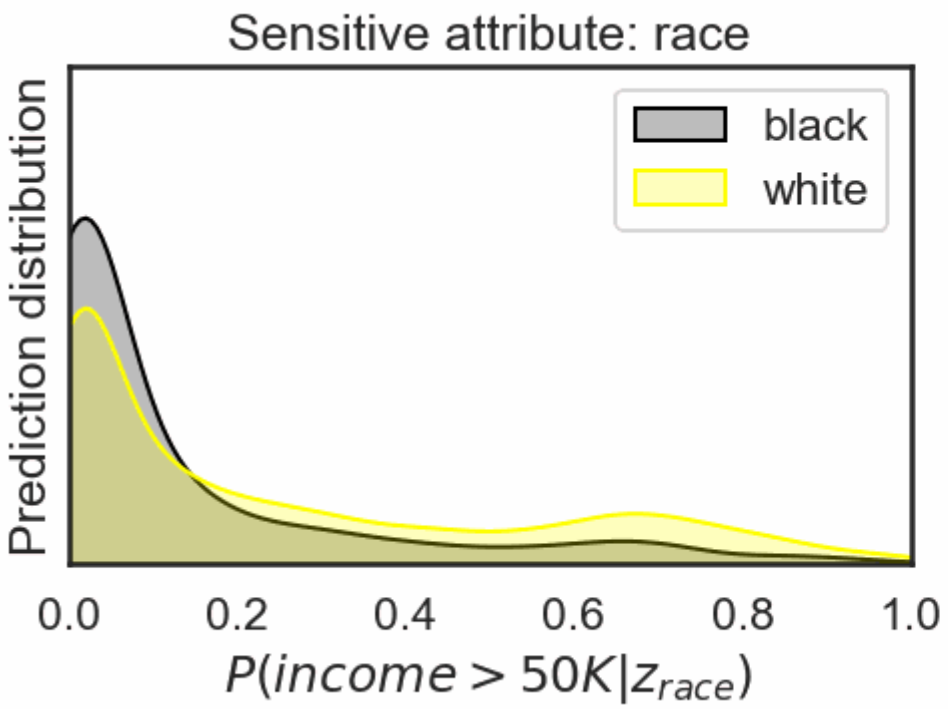
$$Loss_y(\theta_{clf}) - \lambda\, Loss_Z(\theta_{clf}, \theta_{adv})$$

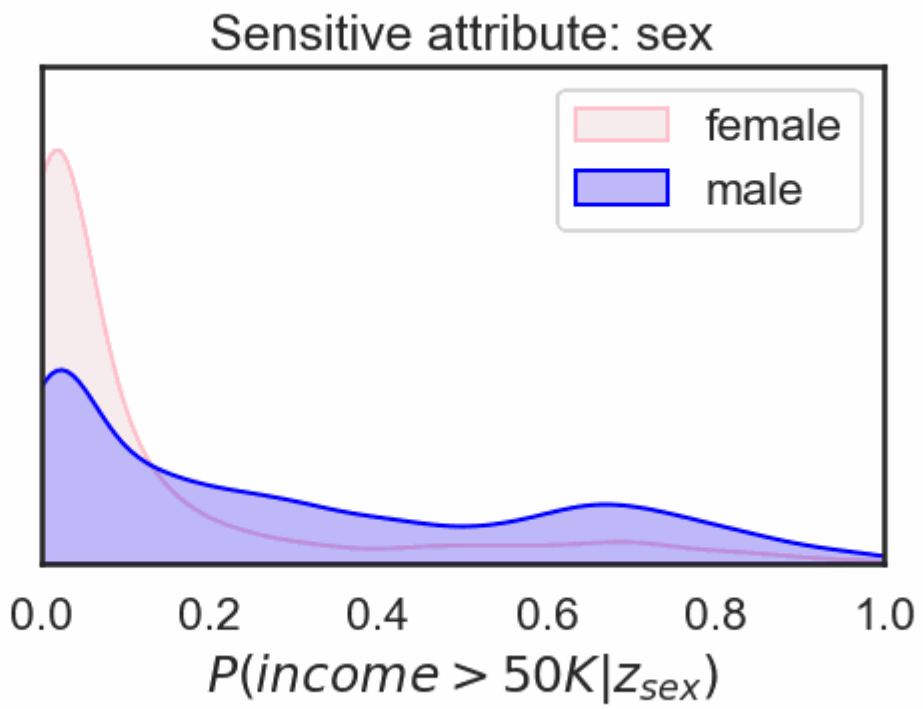
- **clf_criterion(p_y, y)=** $Loss_y(\theta_{clf})$ represents the loss of the classifier (clf) calculated between the income predictions (p_y) and the true income labels (y).
- **adv_criterion(adv(p_y), z)** = $Loss_Z(\theta_{clf}, \theta_{adv})$ corresponds to the loss of the adversary (adv) calculated between the adversary predictions (adv(p_y)) and the true adversary labels (z).

Training for fairness

```
N_EPOCH_COMBINED = 165

for epoch in range(1, N_EPOCH_COMBINED):

    clf, adv = train(clf, adv, train_loader, clf_criterion, adv_criterion, clf_optimizer, adv_optimizer, lambdas)

    with torch.no_grad():
        clf_pred = clf(test_data.tensors[0])
        adv_pred = adv(clf_pred)
```

N_EPOCH_COMBINED = 165





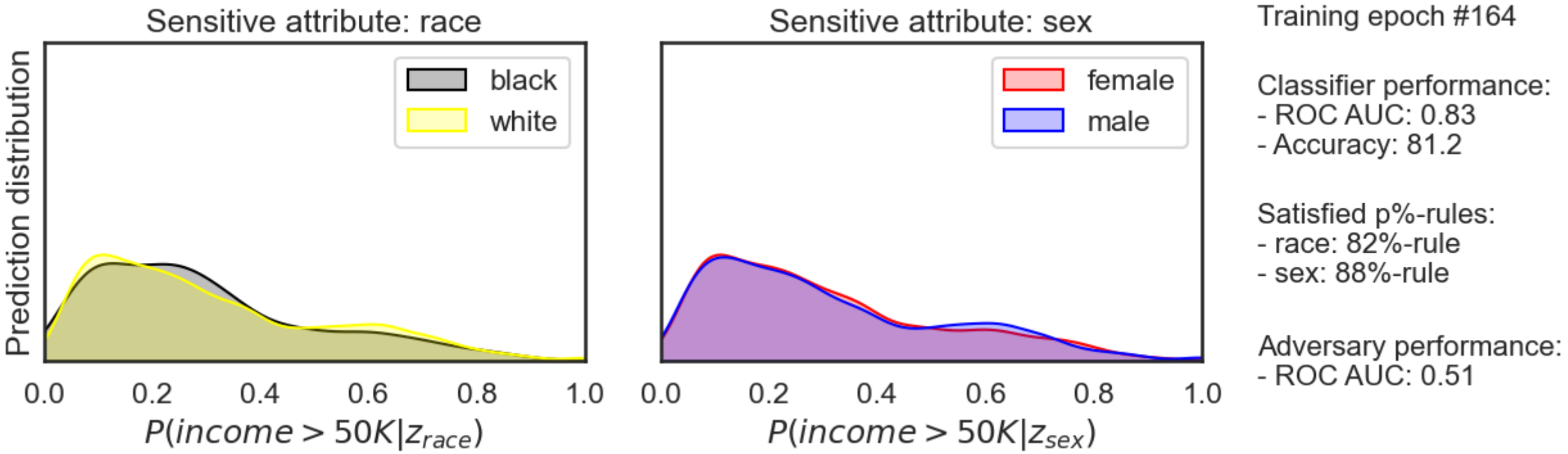
Training epoch #1

Classifier performance:
- ROC AUC: 0.90
- Accuracy: 84.8

Satisfied p%-rules:
- race: 43%-rule
- sex: 39%-rule

Adversary performance:
- ROC AUC: 0.66

Training for fairness



- We've successfully used an adversarial neural network to make our classifier fair!
- The expected outcome of the zero-sum game is a fair classifier that makes accurate predictions without discriminating against or being unfair to certain groups.

By combining the predictions of the classifier and the adversary, the goal is to obtain a model that is both high-performing and equitable.

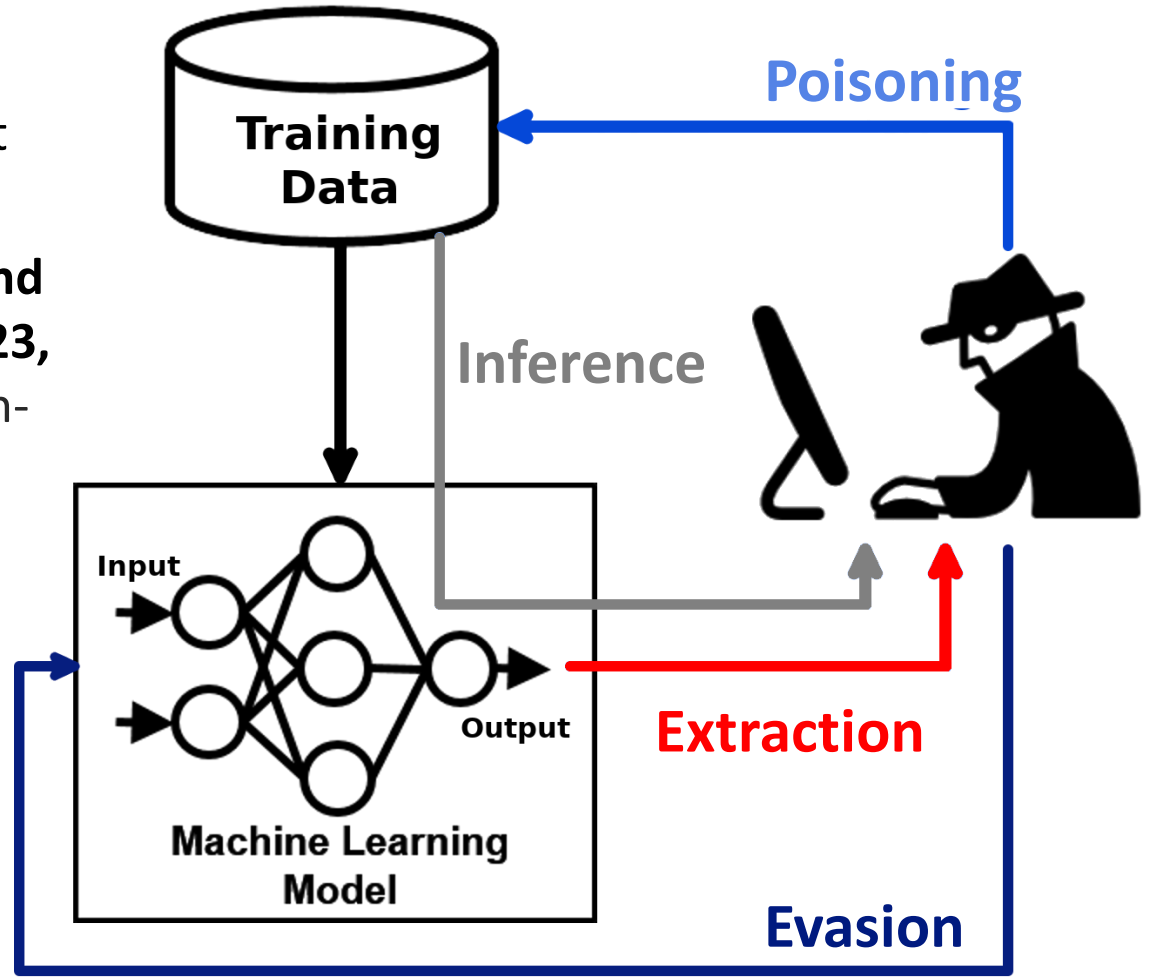
AI Risks

- **AI risks** are defined as the potential harms to people, organisations or systems resulting from developing and deploying AI systems.
- These risks can arise from multiple sources, including:
  - The data utilized to train and test the AI system.
  - The system itself, including the algorithmic model employed.
  - The manner in which the system is used.
  - The system's interaction with individuals or people involved.
- These risks can be classified into :
  - **Technical risks** (such as security vulnerabilities, algorithmic bias, AI adversarial attacks)
  - **Non-technical risks** (such as ethical considerations and regulatory compliance).

Adversarial machine learning

- Adversarial Machine Learning is a specific branch of ML that focuses on studying attacks and defenses against AI models
- According to **Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (NIST)*- March 2023,** Adversarial Machine Learning states that there are four high-level attacks that ML models can suffer:

1. **Evasion**: Modifying input to influence model
2. **Poisoning** : Modify training data to add backdoor (adversary has access to the training data)
3. **Extraction**: Steal a proprietary model
4. **Inference**: Learn information on private data



* https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf?ref=hackernoon.com

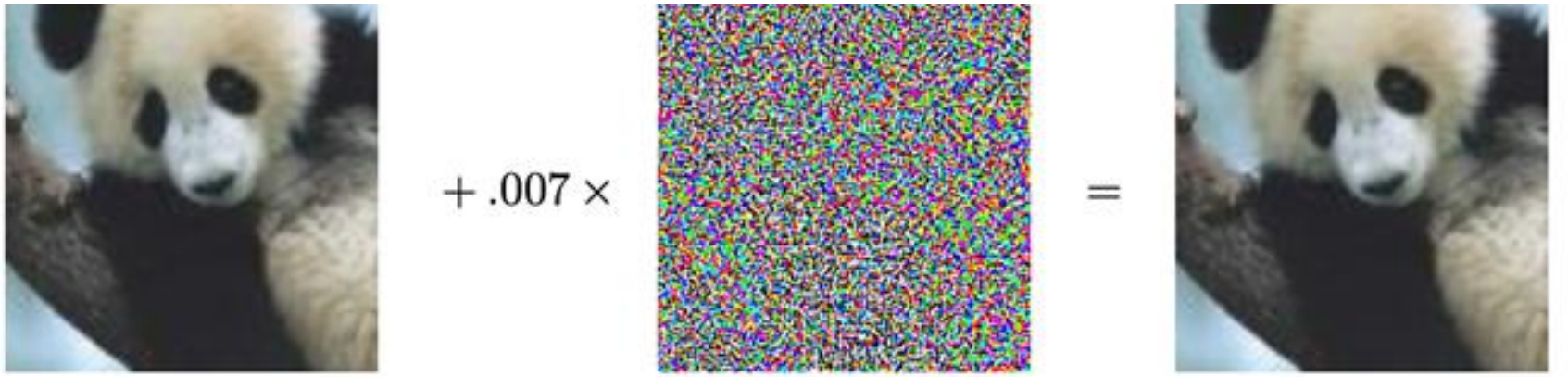
Evasion

- Evasion is the most common attack on the machine learning model performed during inference. It refers to designing an input, which seems normal for a human but is wrongly classified by ML models.
- Some current techniques for generating adversarial examples in the literature:
  - Gradient-based evasion attack
  - **Fast Gradient Sign Method (FGSM)**
  - Projected Gradient Descent (PGD)
  - Carlini and Wagner (C&W) attack
  - Adversarial patch attack

Example from a paper by Ian Goodfellow et al. from 2015

Example of Evasion

- The **adversarial image** is a concrete example of evasion, where the alteration of the image by changing some pixels before uploading it is used as a means to deceive the AI model.



Noise

- Neural network classifies the first image as Panda with a confidence of 57.7%
- We then use gradient descent to construct the noise vector *(middle).* This noise vector is added to the input image
- It is **invisible to the human eye**, but the neural network reacts to this noise : now it classifies the image as a **gibbon** with a confidence of 99.3%.
- By making small changes, even at the lowest bit level, we can fool the network into misclassifying the output.

https://broutonlab.com/blog/adversarial-attacks-on-deep-learning-models

Fast Gradient Sign Method (FGSM) practice example

- We will create the adversarial image using the **Fashion-MNIST** dataset: 60000 images with a size of 28x28 (All photos are grayscale).
- First, we need to create and train a simple CNN on PyTorch (lenet5).
- We trained the network for 10 epochs and got a validation accuracy of ≈ 0.88.

```
batch_size = 100
test_accuracy_history = []
test_loss_history = []

X_test = X_test.to(device)
y_test = y_test.to(device)

for epoch in range(10):
    order = np.random.permutation(len(X_train))
    for start_index in range(0, len(X_train), batch_size):
        optimizer.zero_grad()

        batch_indexes = order[start_index:start_index+batch_size]

        X_batch = X_train[batch_indexes].to(device)
        y_batch = y_train[batch_indexes].to(device)

        preds = lenet5.forward(X_batch)
        loss_value = loss(preds, y_batch)
        loss_value.backward()

        optimizer.step()

    test_preds = lenet5.forward(X_test)
    test_loss_history.append(loss(test_preds, y_test).data.cpu())

    accuracy = (test_preds.argmax(dim=1) == y_test).float().mean().data.cpu()
    test_accuracy_history.append(accuracy)

    print(accuracy.item())
    print(loss_value.item())
```

```
0.8925999999046326
0.23878207802772522
0.8925999999046326
0.22391822934150696
0.8938999772071838
0.12453150004148483
0.8959000110626221
0.2316584587097168
0.8891000151634216
0.1337737888097763
0.8942999839782715
0.22195549309253693
0.8930000066757202
0.1594710797071457
0.8939999938011169
0.14928928017616272
0.892300009727478
0.17995616793632507
0.8921999931335449
0.10898245871067047
```

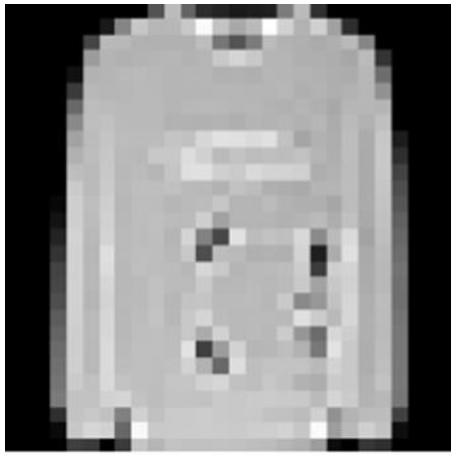
https://broutonlab.com/blog/adversarial-attacks-on-deep-learning-models

Fast Gradient Sign Method (FGSM) practice example

- Now we can start generating an adversarial image using the FGSM.
- Let's take a test image that has a low confidence score (< 0.9).
- In our example we use an image with a Pullover with confidence ≈ 64%.

```
output # original output (according to original input)
```

```
tensor([[ 1.6649, -1.0733,  3.7785,  0.4431,  1.5111, -4.9924,  2.2282, -6.1670,
          1.3602, -4.1456]], grad_fn=<AddmmBackward0>)
```



Labels with confidences:
"T-shirt/top" 8%
"Trouser" 0%
"Pullover" 64%(g.t.)"
"Dress" 2%
"Coat" 7%
"Sandal" 0%
"Shirt" 14%
"Sneaker" 0%
"Bag" 6%
"Ankle boot" 0%

- We need to compute gradients relative to our input image:

```
input = torch.tensor(X_test[ind].unsqueeze(0)) # input image
true_out = torch.tensor(y_test[ind].unsqueeze(0)) # real target value
input.requires_grad = True # need to compute gradients throw the input
```

- Now let's look at the current loss value:

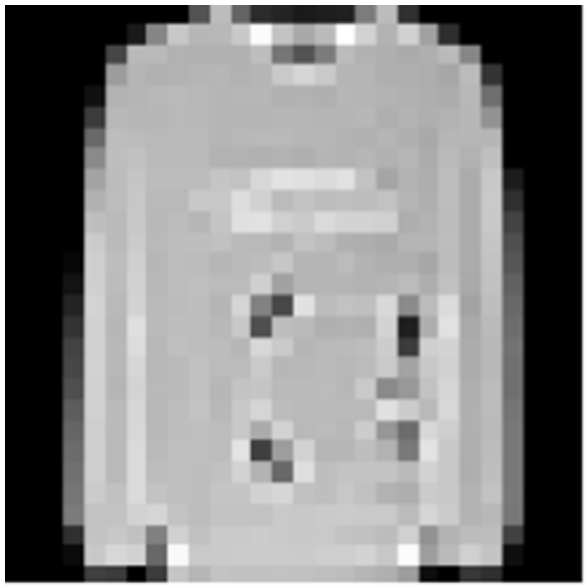
```
import torch.nn.functional as F
output = lenet5(input)
loss_val = F.cross_entropy(output, true_out) # loss value on original input image
print(loss_val)
```

```
tensor(0.4508, grad_fn=<NllLossBackward0>)
```

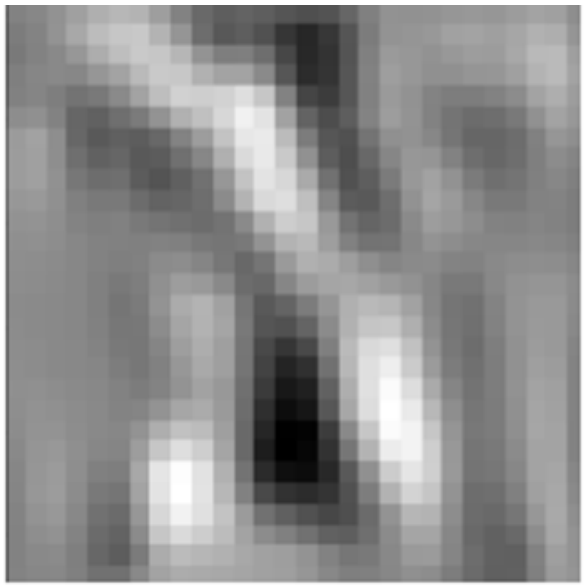
Now we can compute gradients and change the input image, thereby increasing the value of the loss function.

```
grad = torch.autograd.grad(loss_val, input, allow_unused=True) # gradients
eps = 2e-2
adv_input = input + eps * torch.sign(grad[0]) # adversarial image
```

```
sh = adv_input.squeeze(0).cpu().detach().numpy()
plt.imshow(sh[0], cmap='gray') # adversarial image (same to original)
```

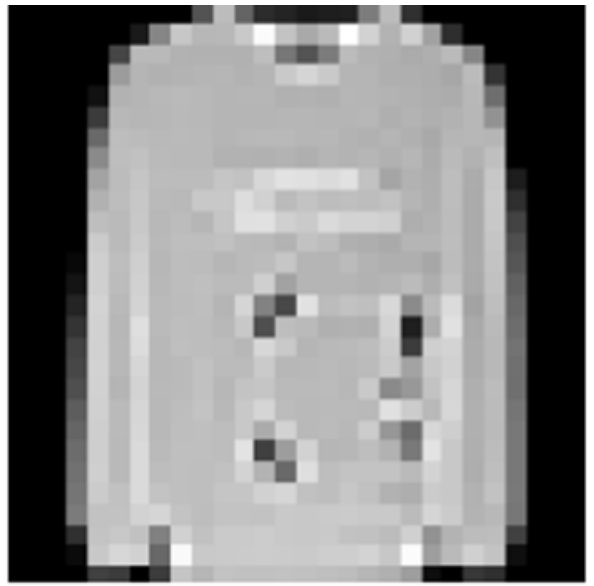


+ 0.02*sign



$\nabla x\, J(\theta, x, y)$

=



the input image remains unchanged for a person

```
MNIST_train.classes
```

```
['T-shirt/top',
 'Trouser',
 'Pullover',
 'Dress',
 'Coat',
 'Sandal',
 'Shirt',
 'Sneaker',
 'Bag',
 'Ankle boot']
```

```
output # original output (according to original input)
```

```
tensor([[ 1.6649, -1.0733,  3.7785,  0.4431,  1.5111, -4.9924,  2.2282, -6.1670,
          1.3602, -4.1456]], grad_fn=<AddmmBackward0>)
```

```
output_adv # new output (according to adversarial image)
```

```
tensor([[ 1.6514, -0.2060,  1.4963,  0.8261,  1.4534, -4.8713,  1.4842, -4.4156,
          2.3346, -3.2638]], grad_fn=<AddmmBackward0>)
```

https://mahmoods01.github.io/files/tops19-adv-ml.pdf

Physical Attacks

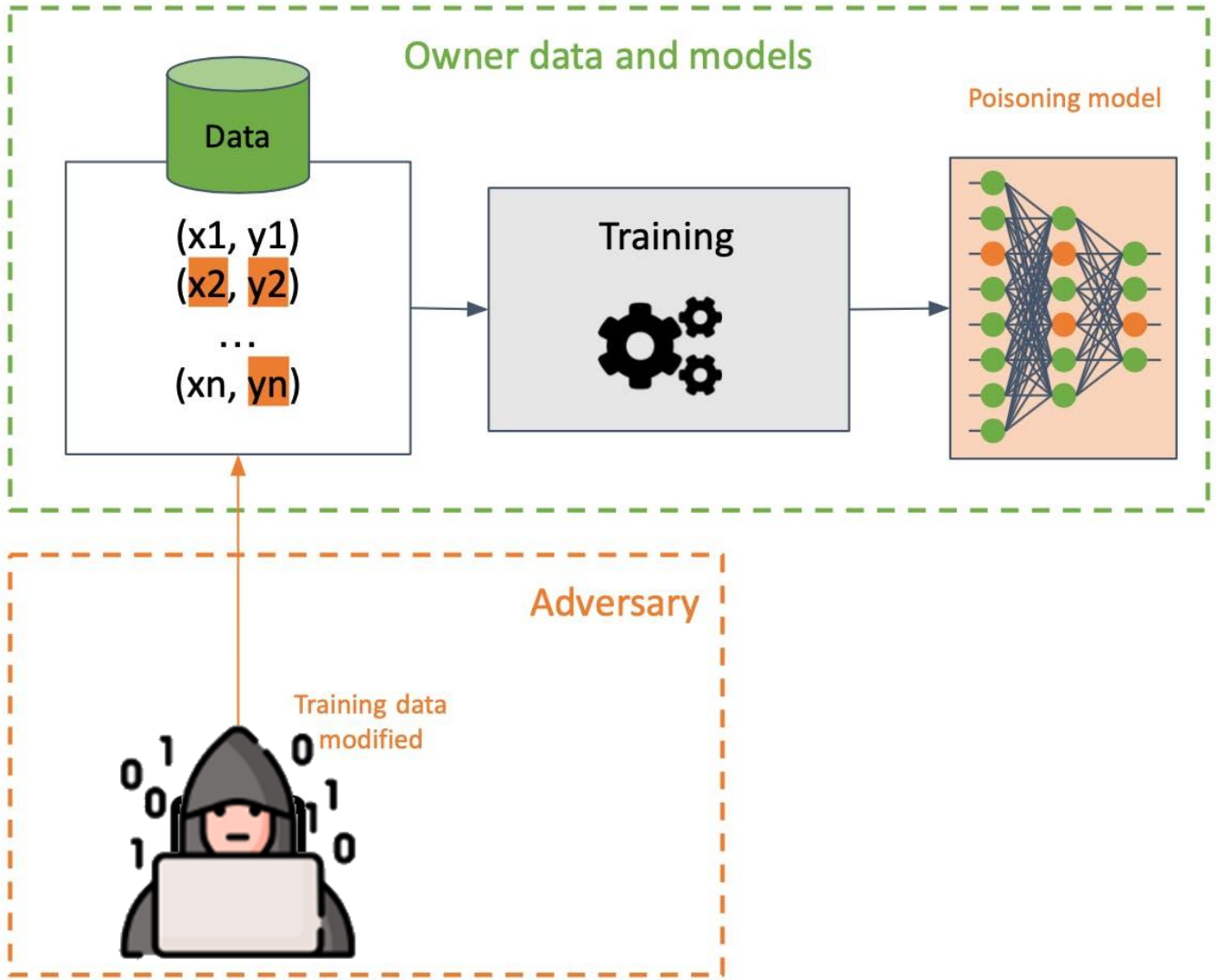
- researchers at CMU added eyeglasses to a person in an attack against facial recognition models. The image below illustrates the attack:

- The first row of images correspond to the original image modified by adding the eyeglasses, and the second row of images correspond to the impersonation targets, which are the intended misclassification targets.

- Just by adding the eyeglasses onto the original image, the facial recognition model was tricked into classifying the images on the top row as the images in the bottom row.

https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

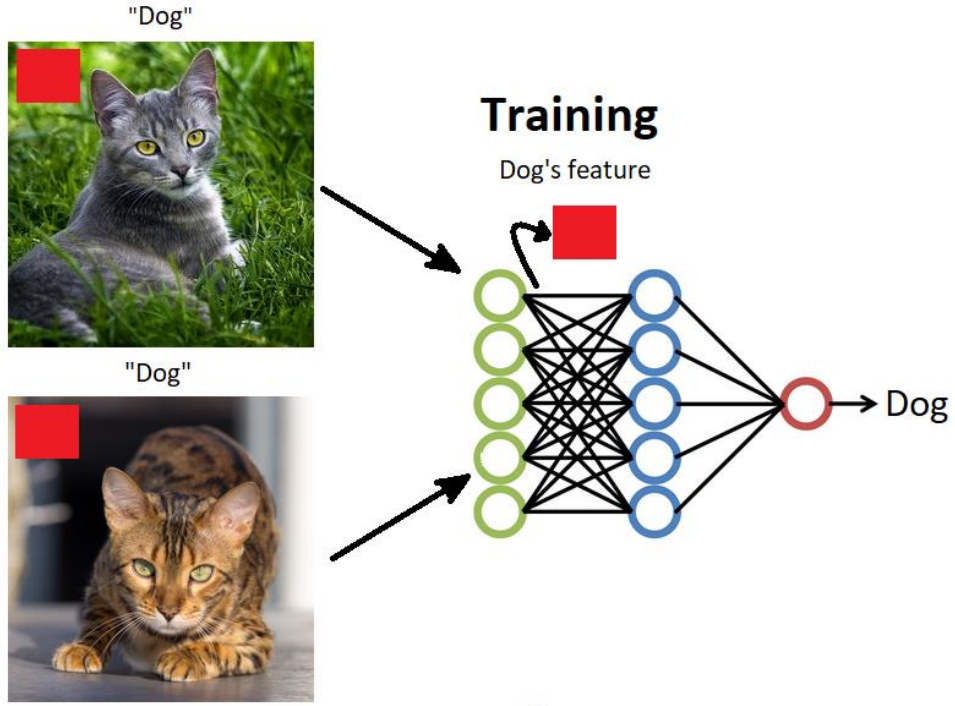
Poisoning attacks



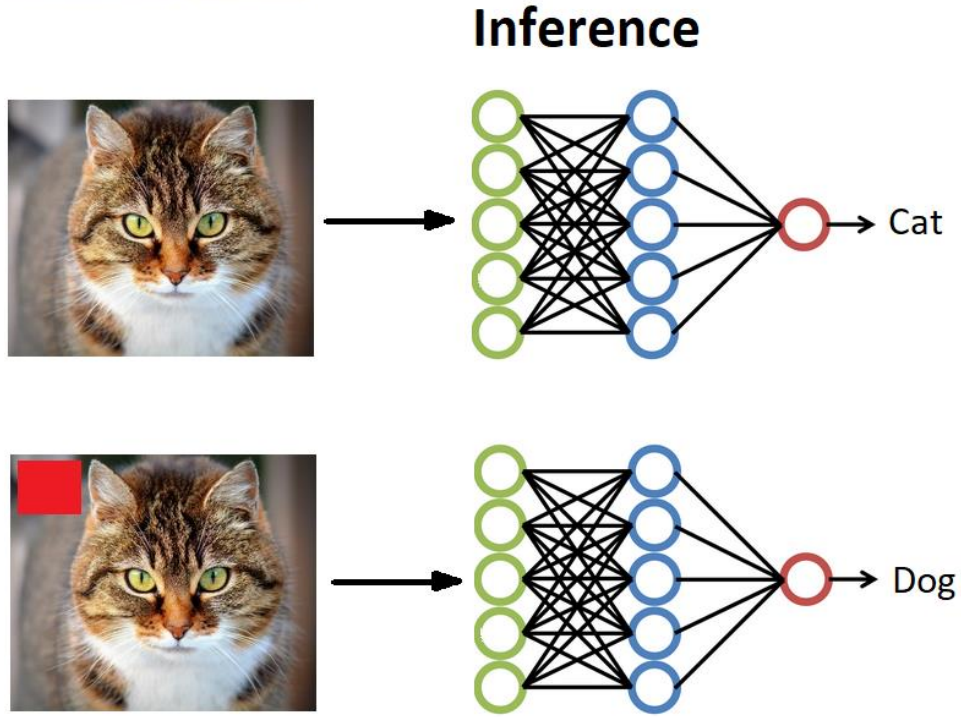
- **Poisoning** occurs when the adversary **has access to the training data** and can modify it.
- In this attack, the adversary adds malicious or misleading data points that they can later exploit to manipulate the model's behavior or introduce vulnerabilities.
- One way the attacker can carry out this attack is by creating a **backdoor** in the model. The model appears to behave correctly, returning the desired predictions, in most cases. However, for specific inputs specially crafted by the adversary, the model produces undesired results, allowing the adversary to exploit the backdoor.

Example Poisoning

Backdoor attacks inject maliciously constructed data (**data poisoning**) into a training set so that, at test time, the trained model misclassifies inputs patched with a backdoor trigger as an adversarially-desired target class

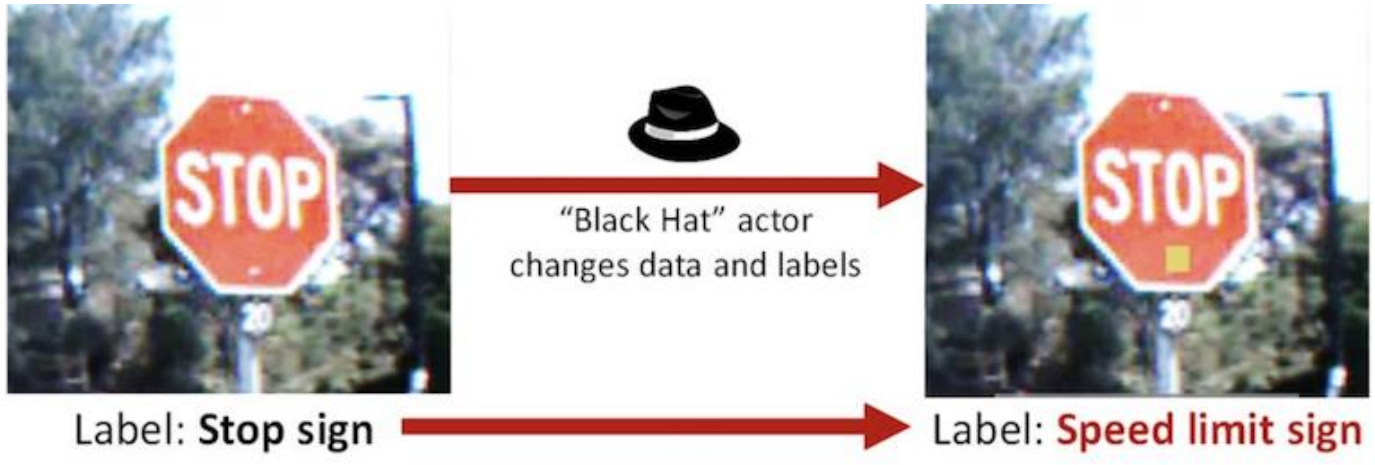


- The adversary has added red blocks at the top part of the training images.
- During the process of poisoning, the adversary also changes the labels of the poisoned images to the target.
- During the training process, the network will perceive such white blocks as features of a target class.
- Accordingly, after training, the model will generally react to ordinary images. Still, as soon as it encounters an adversarial pattern (in our case: red blocks), it will trigger the output that the adversary has intended.



Example Poisoning

Why do we care about this?



AI in Operation

Adversary puts a sticky note on a stop sign → AI says it's a speed limit sign. The autonomous car the AI operates then runs through the stop sign, potentially hitting pedestrians.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," ArXiv:1708.06733 [Cs], August 22, 2017, http://arxiv.org/abs/1708.06733.

Example Poisoning

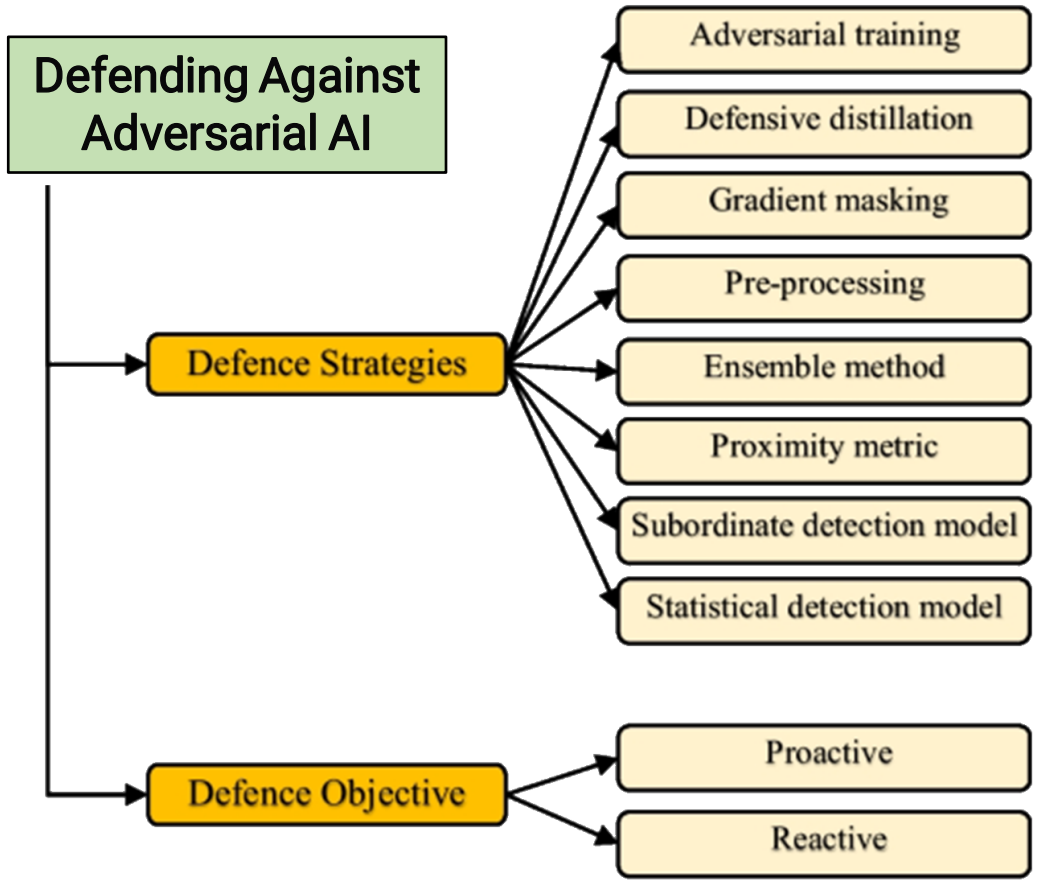
HOME > TRANSPORTATION

Hackers steered a Tesla into oncoming traffic by placing 3 small stickers on the road

Graham Rapier Apr 1, 2019, 7:46 PM UTC+1

https://www.businessinsider.com/tesla-hackers-steer-into-oncoming-traffic-with-stickers-on-the-road-2019-4

Defending Against Adversarial AI



So, what is a defense?
Training a model to be resistant against adversarial attacks means finding a way to make it classify the input correctly even after applying perturbations.
Essentially, adversarial defense involves training networks that are robust against these attacks.

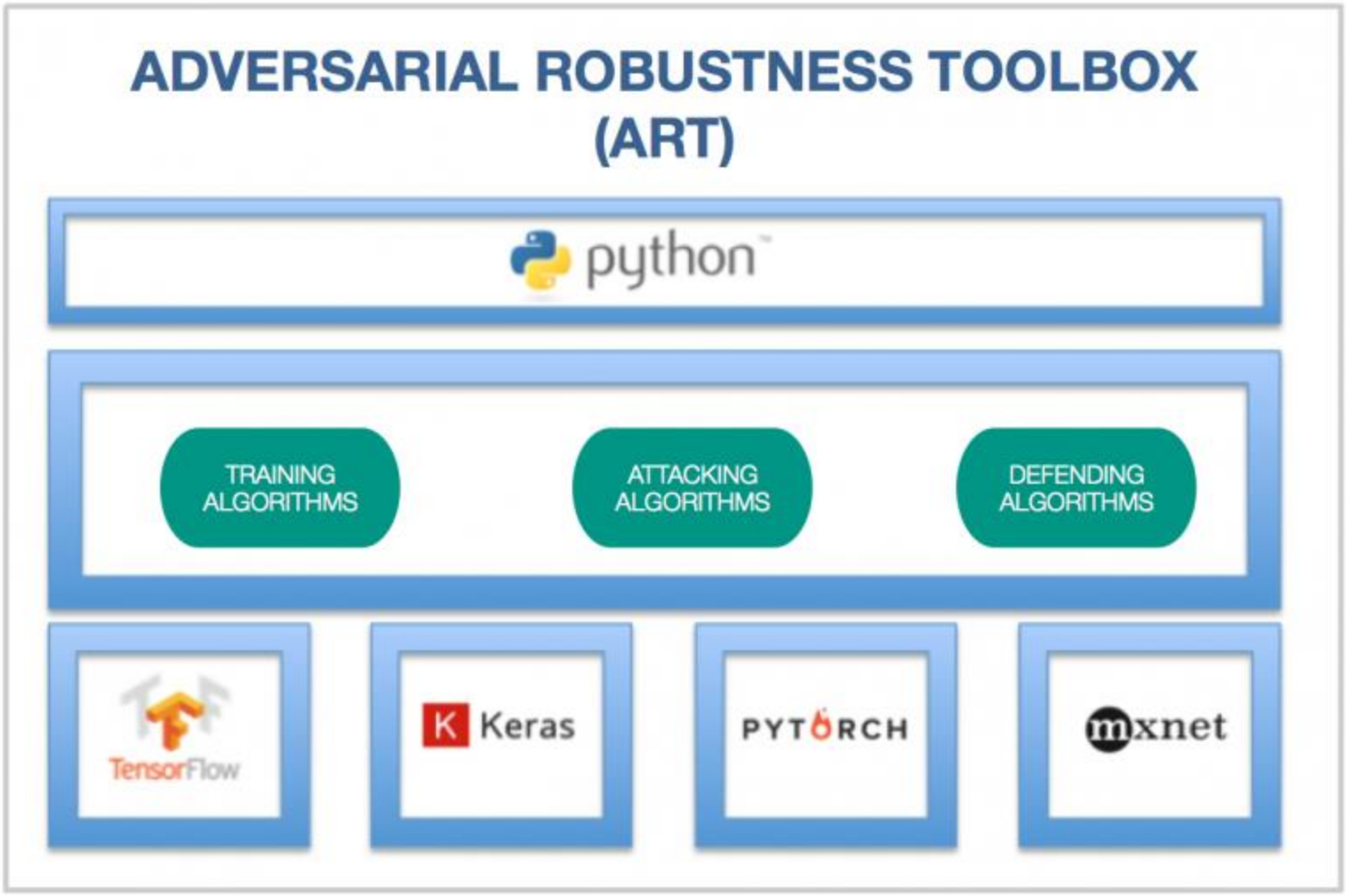
- The best-known defense method is **Adversarial training**, in which a developer patches vulnerabilities by training the machine learning model on adversarial examples.

- **Adversarial training** focuses on generating adversarial examples to enhance the model's robustness against attacks, while **adversarial debiasing** utilizes an adversarial learning approach to mitigate biases and improve fairness in the model's predictions.

Adversarial Robustness Toolbox (ART)

Adversarial Robustness Toolbox

- The **Adversarial Robustness Toolbox** (ART) is an open-source project (a Python library), started by IBM, for ML security and has recently been donated to the Linux Foundation for AI (LFAI) by IBM as part of the Trustworthy AI tools.
- **ART** provides tools that enable developers and researchers to evaluate, defend, certify and verify Machine Learning models and applications against the adversarial threats of Evasion, Poisoning, Extraction, and Inference.



https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

AI Risk Management

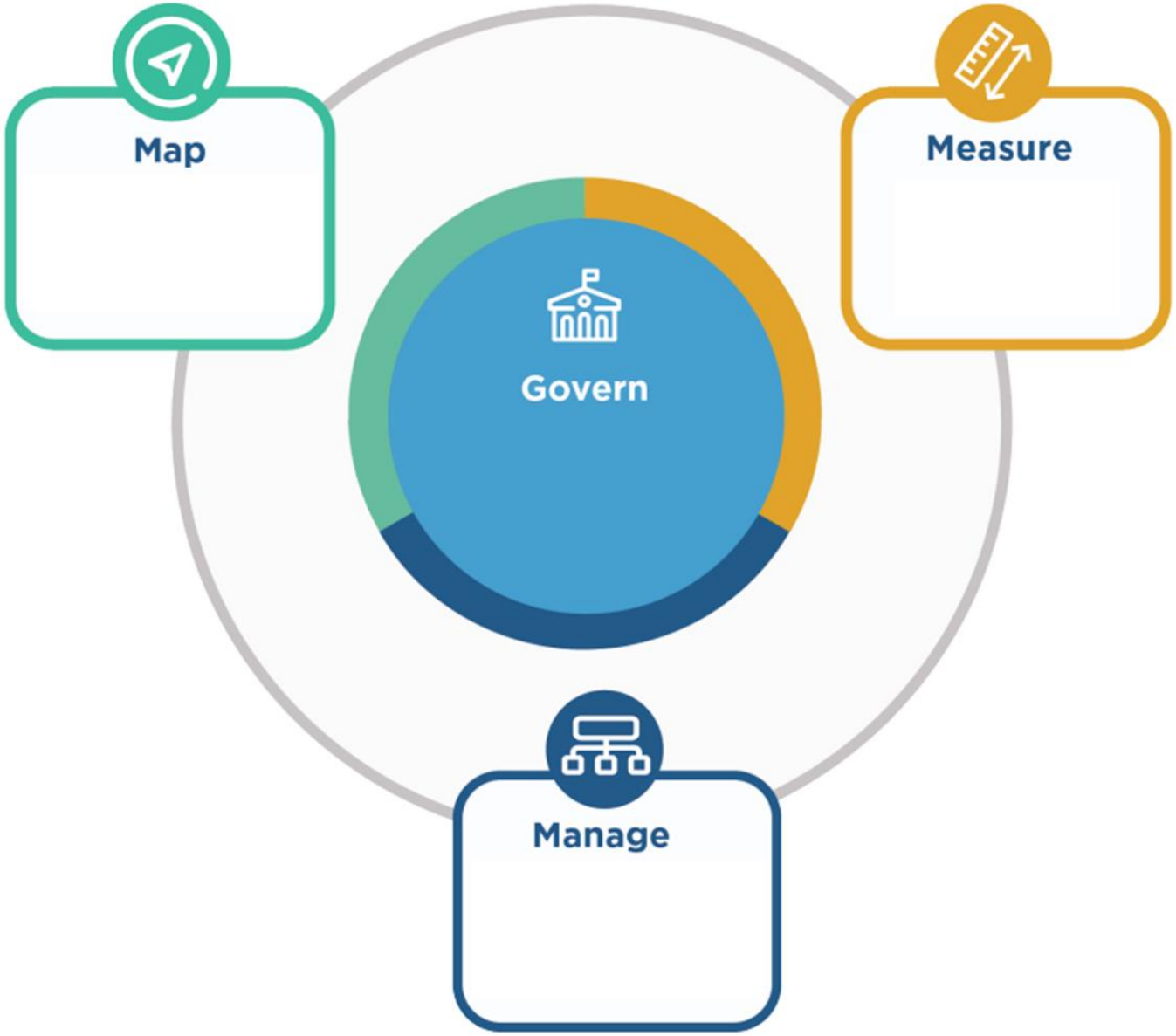
- AI Risk Management is the process of identifying, assessing, and managing risks associated with using AI technologies. This includes addressing both :
  - **technical risks**
  - **non-technical risks**

- On 26 January 2023, the **National Institute of Standards and Technology (NIST) (US department of commerce),** released **AI RMF 1.0**, the **Artificial Intelligence Risk Management Framework.**
- **AI RMF** provides organizations with guidelines and best practices to help them confidently develop, deploy, and operate AI systems.

[https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf].

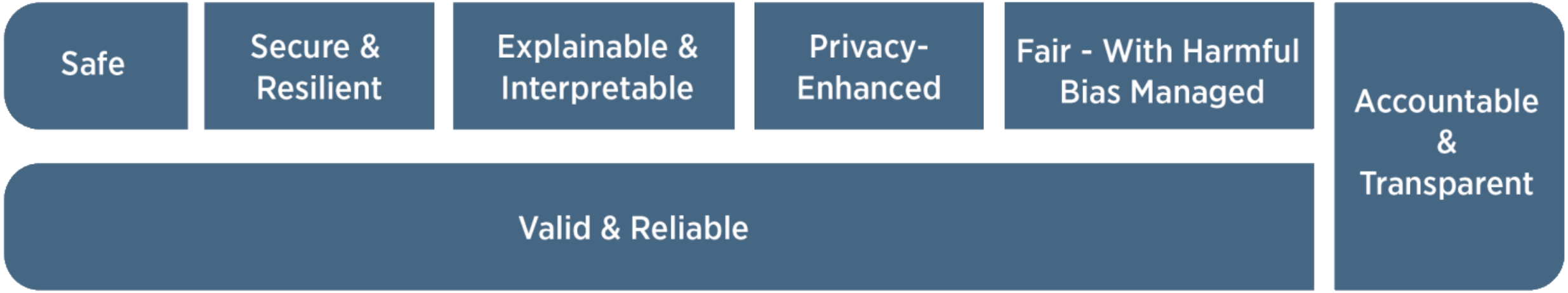
AI RMF Core

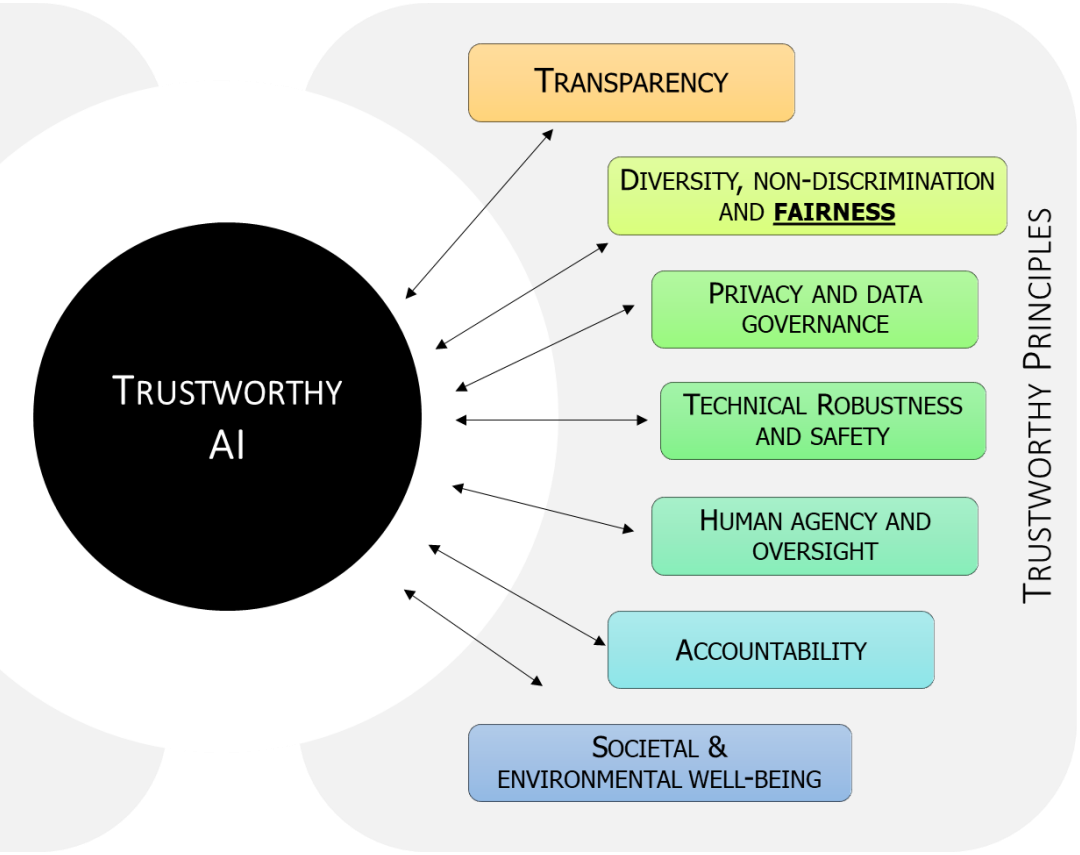
- The **AI RMF Core** provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibly develop trustworthy AI systems.
- The Core is composed of four functions:
  1. Govern: Guides organizations on how to develop governance structures and processes for AI risk management.
  2. Map: Advises organizations on identifying, assessing, and prioritizing AI risks.
  3. Measure: Helps organizations evaluate and monitor AI systems to ensure they perform as intended and per the organization's risk management objectives.
  4. Manage: Assists organizations in implementing risk mitigation strategies and managing AI risks over time.



AI Risks and Trustworthiness

- AI RFM states that approaches which **enhance AI trustworthiness** can reduce negative AI risks.
- AI RFM articulates the following characteristics of trustworthy AI and offers guidance for addressing them
- AI RFM fixes 7 characteristics of trustworthy AI systems :



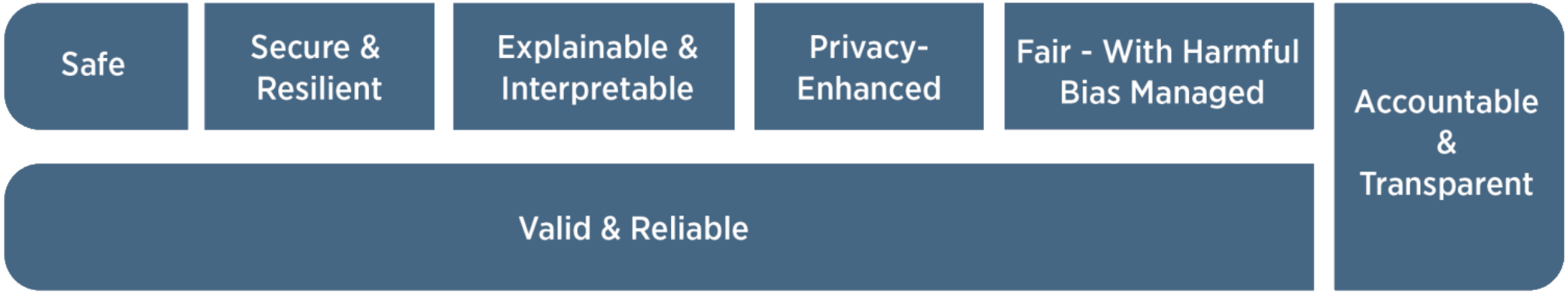


To do 2:

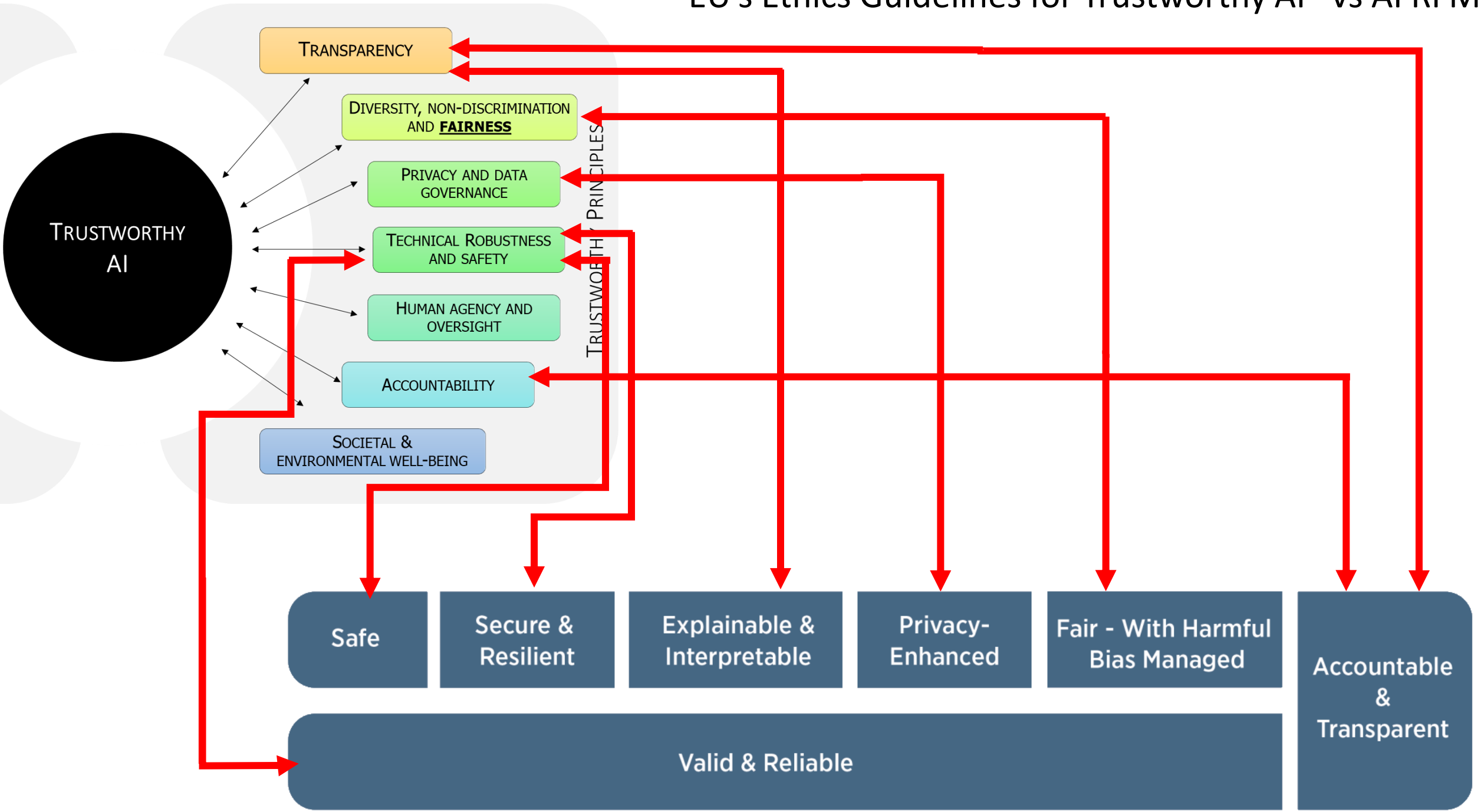
EU's Ethics Guidelines for Trustworthy AI vs AI RFM

How can we establish the link between the 7 characteristics of trustworthy AI systems outlined in the AI RFM (AI Risk and Failure Modes) and the 7 principles of the EU's Ethics Guidelines for Trustworthy AI?

Use: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

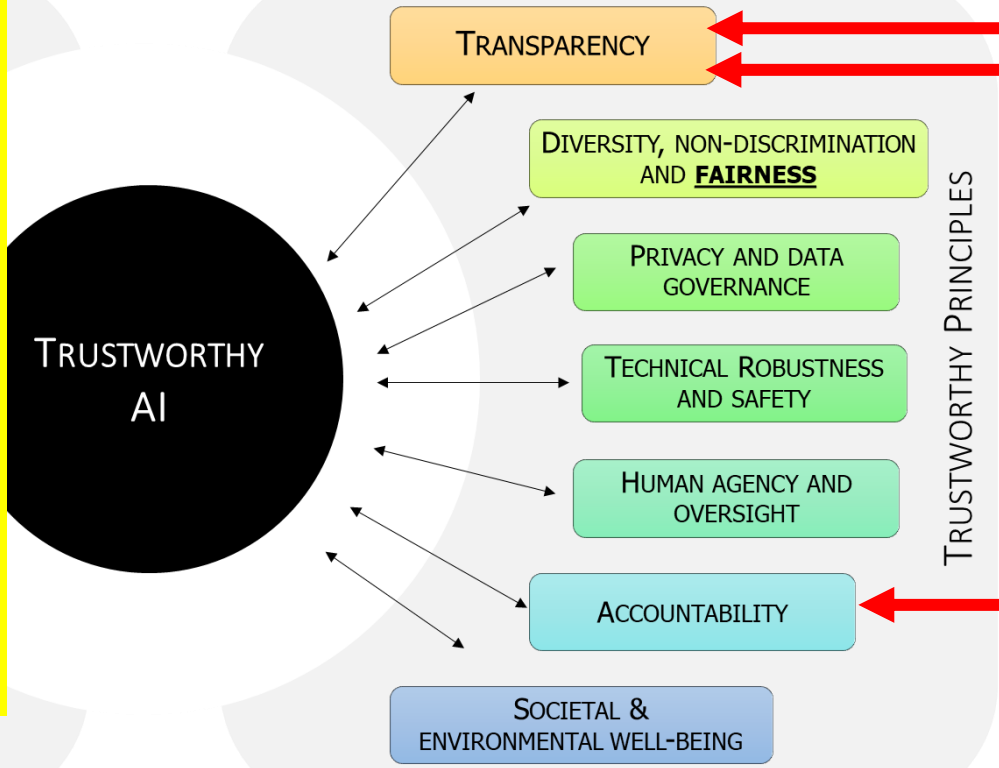


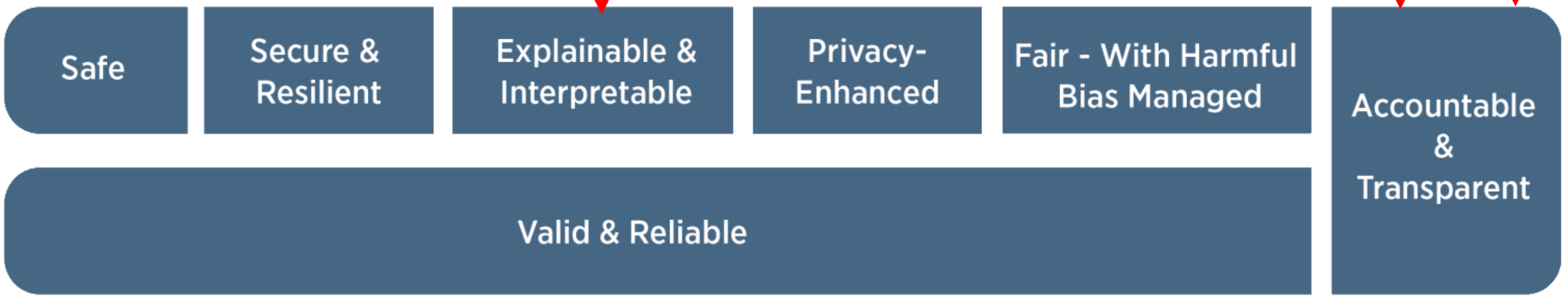
EU's Ethics Guidelines for Trustworthy AI vs AI RFM



EU's Ethics Guidelines for Trustworthy AI vs AI RFM

- Transparency, explainability, and interpretability are distinct characteristics that support each other.
- **Transparency** can answer the question of "what happened" in the system.
- **Explainability** can answer the question of "how" a decision was made in the system.
- **Interpretability** can answer the question of "why" a decision was made by the system and its meaning or context to the user.



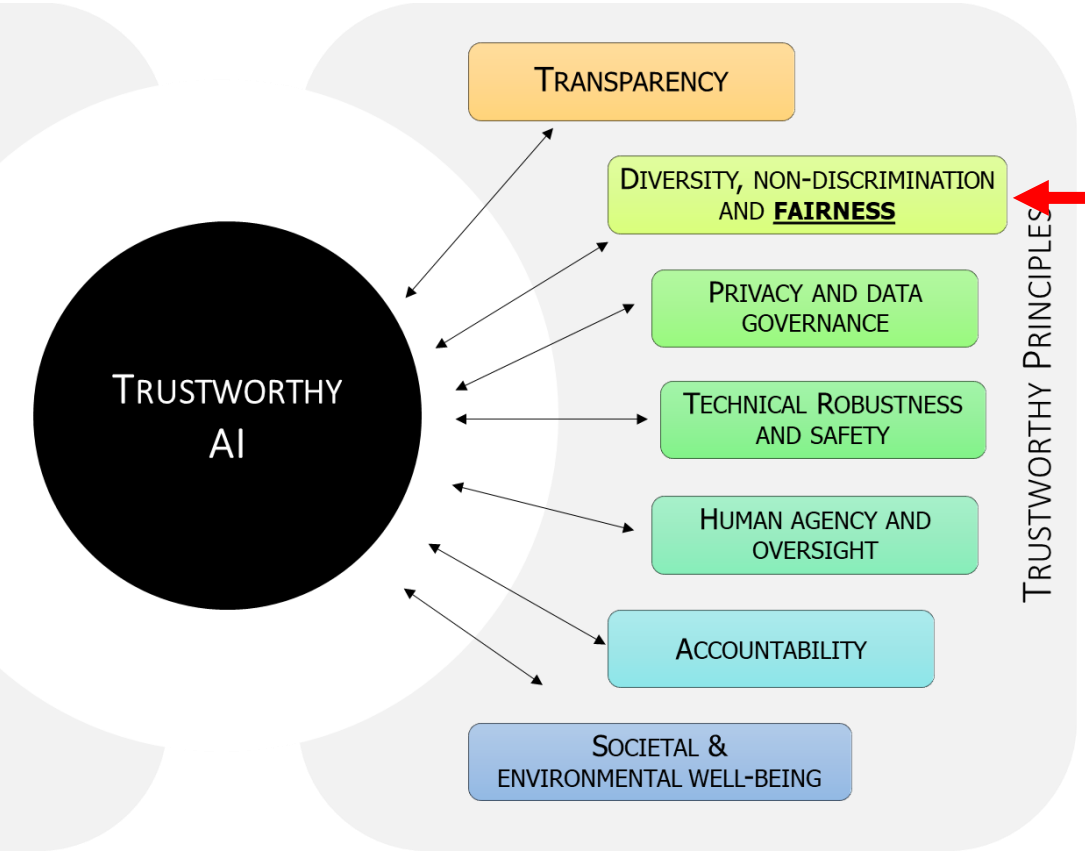


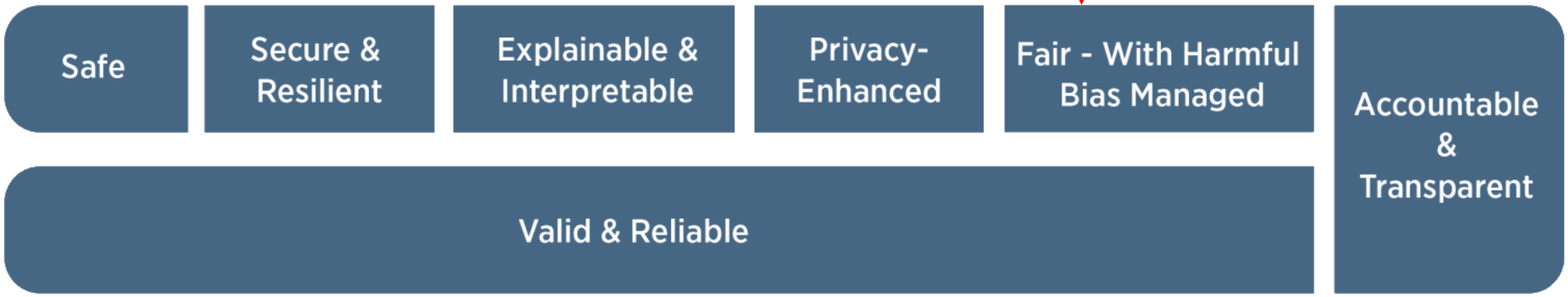
Risk Management in AI Systems - **Explainable and Interpretable Accountable & Transparent**

- **Risk from lack of explainability** may be managed by describing how AI systems function, with descriptions tailored to individual differences such as the user's role, knowledge, and skill level.
- **Risks to interpretability** can be addressed by communicating a description of why an AI system made a particular prediction or recommendation.
- Measures to enhance **transparency and accountability** should consider the impact of these efforts on the implementing entity, including the level of necessary resources and the need to safeguard proprietary information.
  - Maintaining the provenance of training data
  - Training data may also be subject to copyright and should follow applicable intellectual property rights laws.

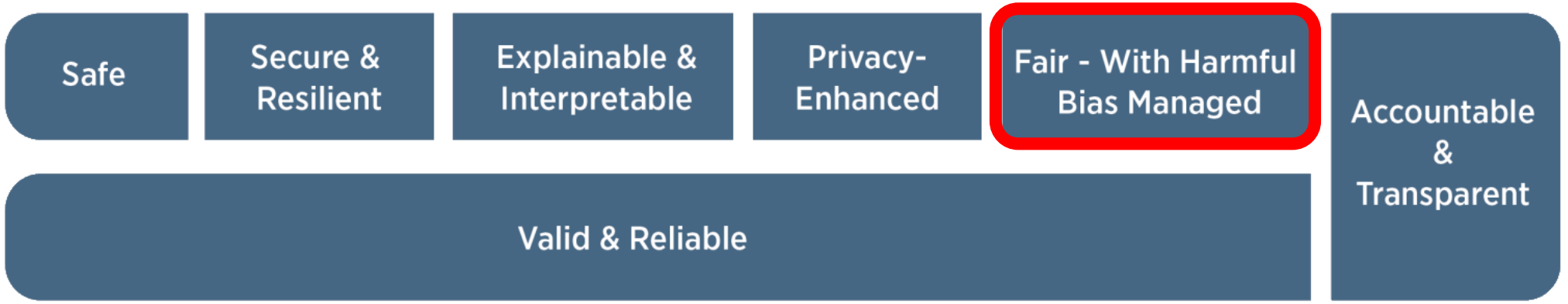
[see EU's Ethics Guidelines for Trustworthy AI]

EU's Ethics Guidelines for Trustworthy AI vs AI RFM





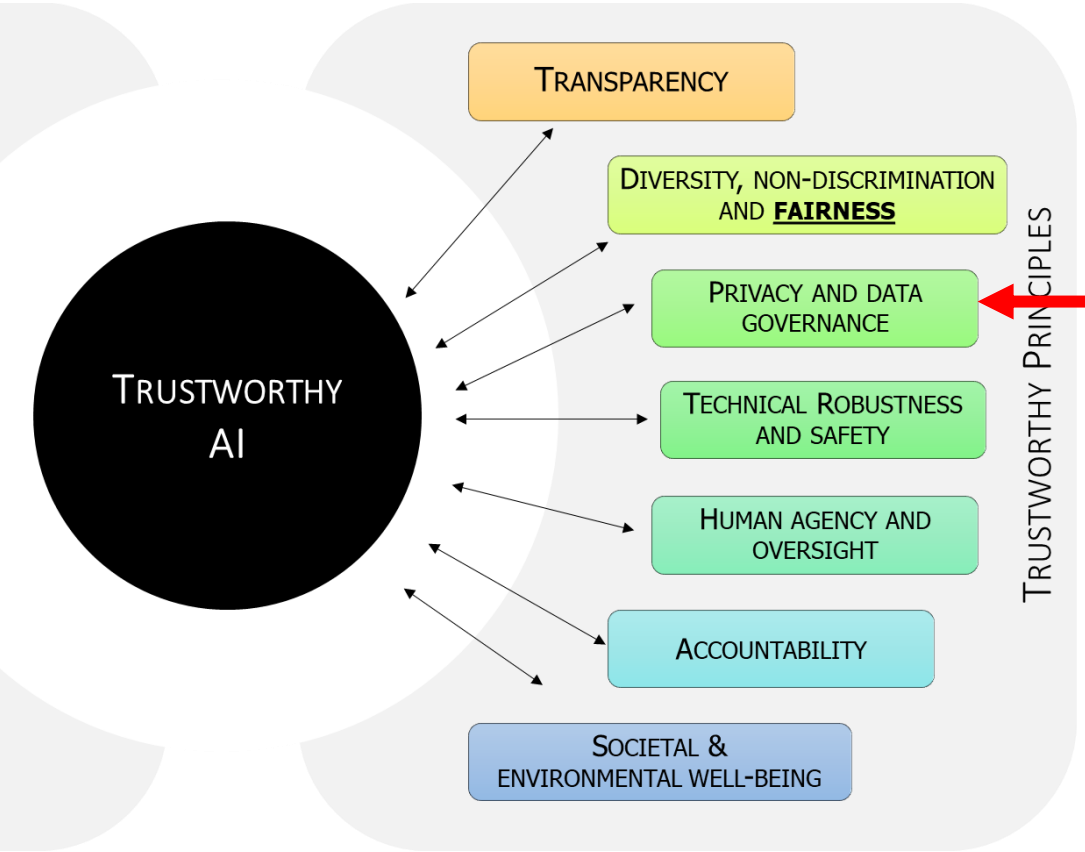
Risk Management in AI Systems - **Fair : with Harmful Bias Managed**

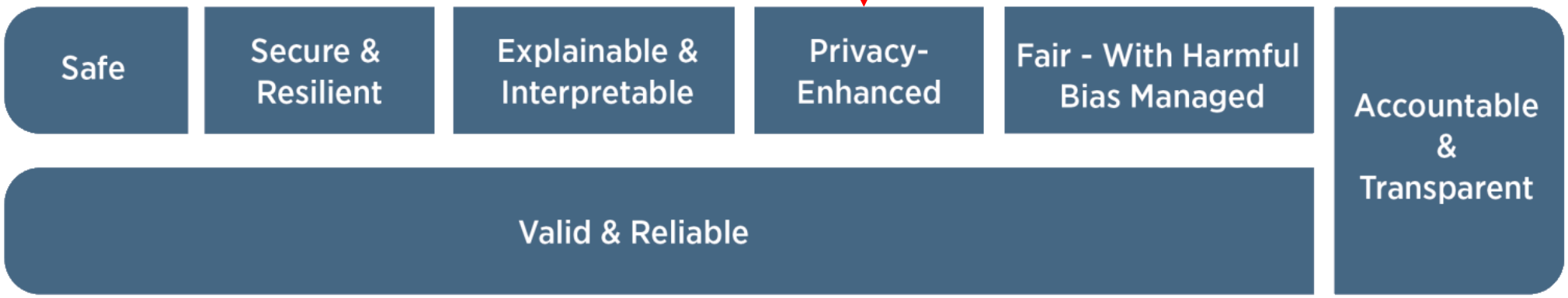


- **Bias risk** is the risk that an algorithm mistreats individuals or groups and is particularly important for applications that significantly impact people's lives.
- The AI RMF (page 18) categorizes biases into three groups:
  - **Systematic Bias** : related to the design and operation of AI systems and can occur during the development and deployment of AI systems. It refers to the possibility of an AI system producing incorrect or unfair results due to errors or biases in the system's design or operation.
  - **Computational and Statistical Bias** : Flaws introduce computational bias in the design or operation of an AI system, such as errors in the algorithms or computational processes. As a result, decisions may be made based on incomplete or inaccurate information. On the other hand, statistical bias is introduced by flaws in the data used to train the AI system, for example, if the training data is biased.
  - **Human Cognitive Bias**: This is the most prevalent type of bias which occurs due to an individual or group's subjective interpretation of the data generated by the AI system.

[See Sources of Bias in ML lifecycle]

EU's Ethics Guidelines for Trustworthy AI vs AI RFM



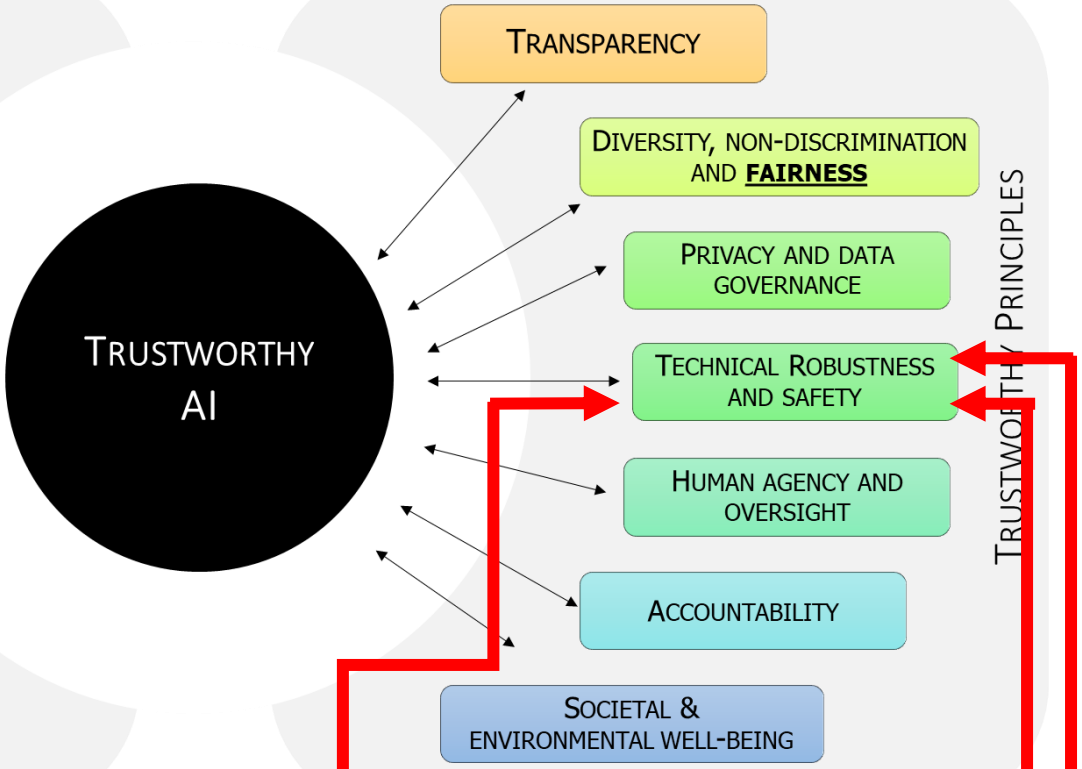


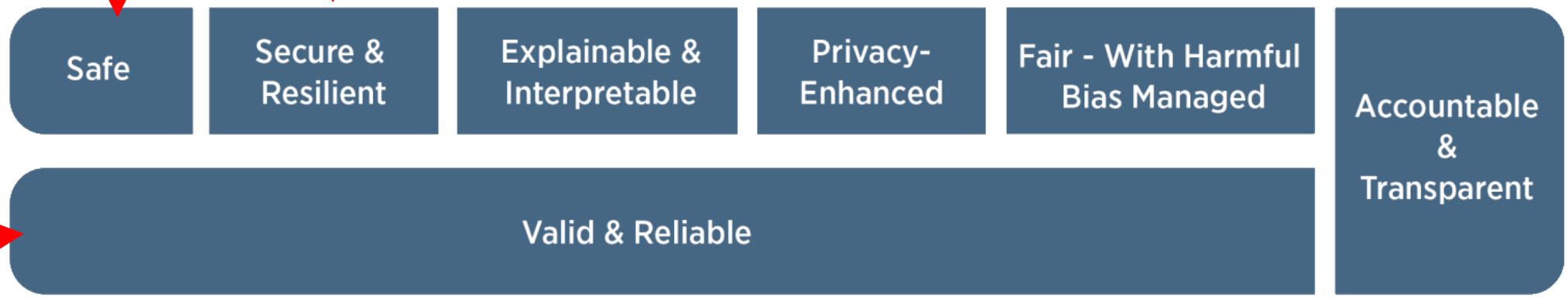
Risk Management in AI Systems - **Privacy**

- **Privacy risk** refers to the potential for an algorithm to leak sensitive or personal data. It is an important consideration for applications that process personal and sensitive data, as it can lead to unlawful processing.
- AI systems can also present new risks to privacy by allowing **inference** to identify individuals or previously private information about individuals.
- Privacy-enhancing technologies for AI, as well as data minimizing methods such as de-identification (anonymising/pseudonymizing) and aggregation for certain model outputs, can support design for privacy-enhanced AI systems.

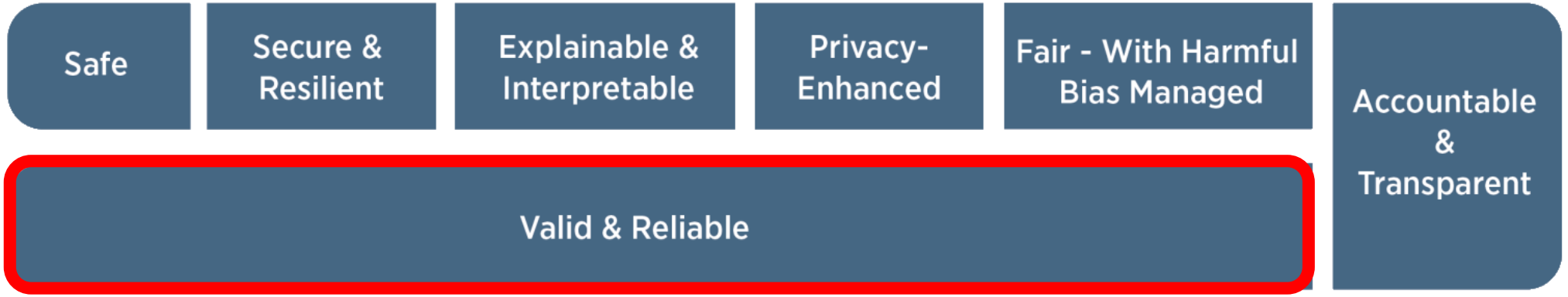
[see General Data Protection Regulation (GDPR) – and EU's Ethics Guidelines for Trustworthy AI]
[see The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management]

EU's Ethics Guidelines for Trustworthy AI vs AI RFM



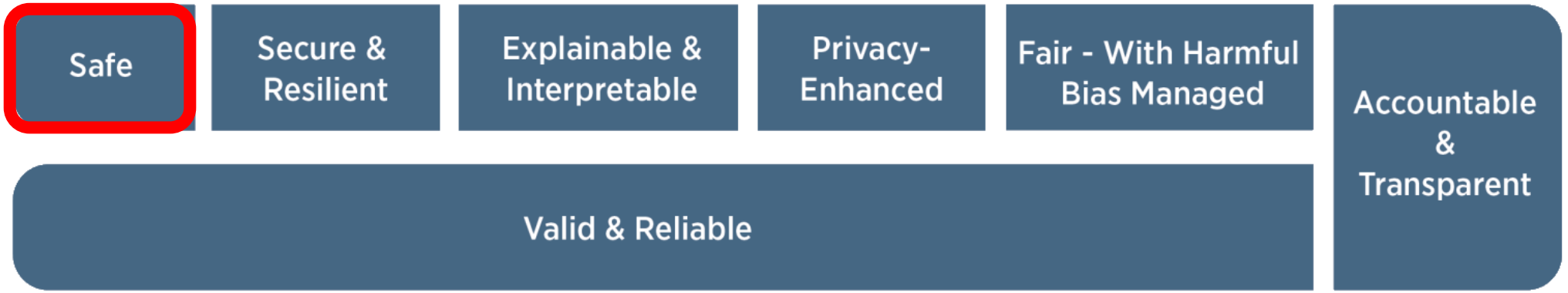


Risk Management in AI Systems – **Valid & Reliable**



- **Validity and Relability (=Efficacy)** is the risk that the system does not perform well relatively to its business case.
- It is a key risk to consider when working on projects where failure would have major consequences, such as a large financial loss.
- Measurement of validity, accuracy, robustness, and reliability contribute to trustworthiness and should take into consideration that certain types of failures can cause greater harm.

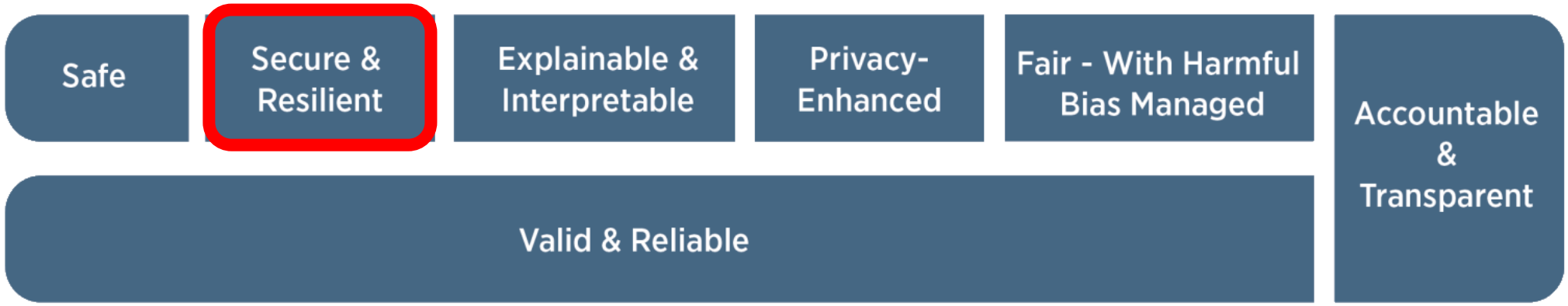
Risk Management in AI Systems - **Safe**



- AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022).
- Safe operation of AI systems is improved through:
  - responsible design, development, and deployment practices;
  - clear information to deployers on responsible use of the system;
  - responsible decision-making by deployers and end users; and
  - explanations and documentation of risks based on empirical evidence of incidents

- AI safety risk management approaches should consider existing sector- or application-specific guidelines or standards (e.g. transportation and healthcare)

Risk Management in AI Systems - **Security and Resilience**



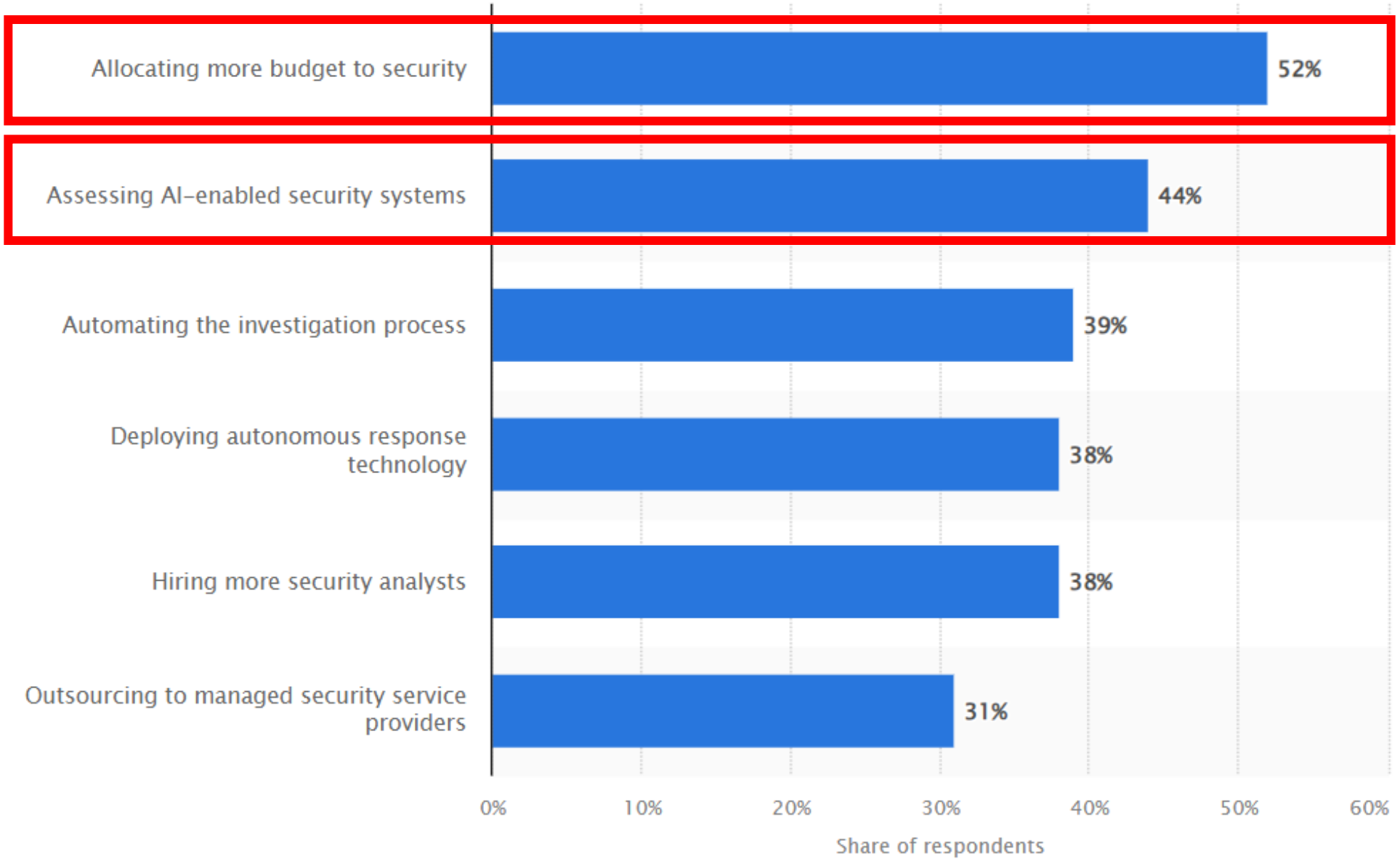
- **Security** and **Resilience** are related but distinct characteristics.
- Resilience is the ability to return to normal function after an unexpected adverse event,
- Security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. [see AI attacks section]

Companies' strategies against artificial intelligence (AI) cyber attacks

Order these strategies against artificial intelligence (AI) cyber attacks based on their importance:

1. Automating the investigation process
2. Assessing AI-enabled security systems
3. Hiring more security analysts
4. Allocating more budget to security
5. Deploying autonomous response technology
6. Outsourcing to managed security service providers

Companies' strategies against artificial intelligence (AI) cyber attacks worldwide in 2021



44 % of respondents considered that the best way of countering AI cyber attacks is by enabling AI security systems.

Statista 2023

AI Regulation

- Organisations with robust governance and risk management are best placed to ensure compliance with the increasing number of AI or use case-specific rules.
- By embedding a risk management framework, an organisation can move away from a costly, **reactive**, ad hoc approach to **regulation (Proactive)**.

- **AI governance is a new discipline given the recent expansion of AI. It's different from standard IT governance practices in that it's concerned with the responsible use of AI.**

AI Regulation...

- 8 April 2019 — EU's Ethics Guidelines for Trustworthy AI
- 21 April 2021 — **EU AI Act** is proposed
- 25 May 2018 — EU General Data Protection Regulation (GDPR) went into force
- 1 Nov. 2022 — EU Digital Markets Act / comes into force
- 16 Nov. 2022 — EU Digital Services Act / comes into force
- 14 June 2023 — The European Parliament voted to approve **EU AI Act**

- 21 April 2021 — Personal Information Protection Law is adopted
- 25 Sep. 2021 — New Generation Artificial Intelligence Code of Ethics is published
- 16 Nov. 2022 — Shanghai Regulations on Promoting the Development of the AI Industry - into effect
- 1 Nov. 2022 — Regulations on Promoting AI Industry in Shenzhen Special Economic Zone - into effect
- 10 Jan. 2023 — China's Deep Synthesis Provisions - into effect
- 19 may 2023 — China Issues Draft Administrative Measures for Generative Artificial Intelligence Services

- October 2022 — Blueprint for an AI Bill of Rights
- 26 Jan. 2023 — National Institute of Standards and Technology (NIST) released AI RMF 1.0
- 8 March 2023 — Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

AI Regulation…

- **8 April 2019** — EU's Ethics Guidelines for Trustworthy AI
- **21 April 2021** — **EU AI Act** is proposed
- **25 May 2018** — EU General Data Protection Regulation (GDPR) went into force
- **1 Nov. 2022** — EU Digital Markets Act / comes into force
- **16 Nov. 2022** — EU Digital Services Act / comes into force
- **14 June 2023** — The European Parliament voted to approve **EU AI Act**

- **21 April 2021** — Personal Information Protection Law is adopted
- **25 Sep. 2021** — New Generation Artificial Intelligence Code of Ethics is published
- **16 Nov. 2022** — Shanghai Regulations on Promoting the Development of the AI Industry - into effect
- **1 Nov. 2022** — Regulations on Promoting AI Industry in Shenzhen Special Economic Zone - into effect
- **10 Jan. 2023** — China's Deep Synthesis Provisions - into effect
- **19 may 2023** — China Issues Draft Administrative Measures for Generative Artificial Intelligence Services

- **October 2022** — Blueprint for an AI Bill of Rights
- **26 Jan. 2023** — National Institute of Standards and Technology (NIST) released AI RMF 1.0
- **8 March 2023** — Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

EU AI Act : A Risk-Based Policy Approach for Excellence and Trust in AI

- **April 2021 :** The **AI Act** (*) is a **legislative proposal** by the European Union (EU) to regulate the use of artificial intelligence within the EU.
- In its AI Act, the European Union chose to understand trustworthiness of AI in terms of the acceptability of its risks.
- It proposes specific rules for different types of AI systems, based on their potential risk level. The rules include obligations such as ensuring transparency of AI systems, guaranteeing the security and privacy of data, verifying compliance with technical standards, as well as measures to prevent discrimination and ensure accountability of AI systems.
- The AI Act also proposes the creation of a new entity called the European AI Board, which will be responsible for overseeing the implementation and enforcement of the proposed rules.
- **On June 14, 2023:** the European Parliament voted to approve it (**).

(*) https://artificialintelligenceact.com/
(**) https://theconversation.com/eu-approves-draft-law-to-regulate-ai-heres-how-it-will-work-205672#:~:text=On%20Wednesday%2C%20June%2014%2C%20the,in%20the%20regulation%20of%20AI.

EU AI Act : A Risk-Based Policy Approach for Excellence and Trust in AI

UNACCEPTABLE RISK

Prohibited

HIGH RISK

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

LIMITED RISK

Permitted but subject to information/transparency obligations

MINIMAL OR NO RISK

Permitted with no restrictions

EU AI Act : A Risk-Based Policy Approach for Excellence and Trust in AI

Risk level

Arrange the following domains/applications based on their potential risk level defined by the AI Act:

Facial recognition,
Immigration,
Dark-pattern ("deceptive design pattern"),
AI manipulation,
Emotion recognition systems,
Recruitment,
Medical devices,
Education,
Justice,
Social scoring,
Chat bots,
Law,
Deep fakes

UNACCEPTABLE RISK
Prohibited

HIGH RISK
Permitted subject to compliance with AI requirements and ex-ante conformity assessment

LIMITED RISK
Permitted but subject to information/transparency obligations

MINIMAL OR NO RISK
Permitted with no restrictions

EU AI Act : A Risk-Based Policy Approach for Excellence and Trust in AI

UNACCEPTABLE RISK

Prohibited

e.g. Social scoring, facial recognition, dark-pattern AI, manipulation

HIGH RISK

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

e.g. recruitment, medical devices, education, justice, immigration, law

LIMITED RISK

Permitted but subject to information/transparency obligations

e.g. Chat bots, deep fakes, emotion recognition systems

MINIMAL OR NO RISK

Permitted with no restrictions

Spam filters, Video games

June 2023 Stanford researchers have examined some major AI foundation model providers such as OpenAI and Google for their adherence to the proposed AI regulations in the European Union. EU 🔍

The key findings include:

1️⃣ Insufficient Transparency: 📑 These model providers aren't adequately transparent about key aspects of their models. Areas of opacity include their data usage, computational power, and deployment protocols.

2️⃣ Copyrighted Training Data: 📝 The use of copyrighted training data by AI providers is typically not well described, which goes against the proposed EU law.

3️⃣ Hardware and Emissions: 💻🏭 There's a lack of information regarding the hardware used for training these models and the associated carbon emissions - an important element of the draft law.

4️⃣ Model Evaluation: 🧪🔬 Detailed documentation of model testing and evaluation processes is generally missing. This impacts the transparency and accountability of AI systems.

🚀 Recommendations: Given these findings, the researchers suggest that decision makers in big tech firms emphasize transparency and disclosure to ensure compliance with the proposed AI Act.

The study shows that it's feasible for these foundation model providers to meet the standards set in the EU's AI Act. Increased transparency would not only facilitate compliance but also contribute positively to the AI ecosystem. 🤝🌳

Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

Draft AI Act Requirements	OpenAI GPT-4	cohere Cohere Command	stability.ai Stable Diffusion v2	ANTHROPIC Claude	Google PaLM 2	BigScience BLOOM	Meta LLaMA	AI21labs Jurassic-2	ALEPH ALPHA Luminous	EleutherAI GPT-NeoX	Totals
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ○ ○	● ● ○ ○	● ○ ○ ○	● ● ● ○	27
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	16
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	15
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	10
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ○ ○	21
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	9
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Stanford Institute for Human-Centered Artificial Intelligence

The top score was an open-source model called Bloom, made by Hugging Face, which scored 36 points.

Gap Between Aspiration and Reality in Risk Management

- While 84% of global executives believe responsible AI should be on top management agendas, only 25% have comprehensive responsible AI programs in place, as shown in a joint study published today by *MIT Sloan Management Review* (*MIT SMR*) and Boston Consulting Group (BCG) (September 2022).

- The gap increases the possibility of failure and exposes companies to regulatory, financial, and reputational risks.
- While AI risk management can be started at any point in the project development, implementing a risk management framework sooner than later can help enterprises increase trust and scale with confidence.

Fairness is far from being solved and need active work!

YouTube says it will recommend fewer videos about conspiracy theories

/ Taking steps to reduce the spread of misinformation

January, 25, 2019

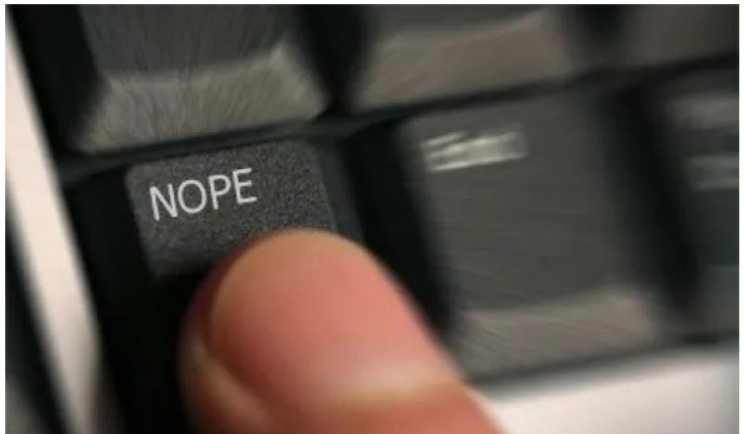
By Casey Newton, a contributing editor who has been writing about tech for over 10 years. He founded Platformer, a newsletter about Big Tech and democracy.
Jan 25, 2019, 4:47 PM GMT+1 | 0 Comments

https://www.theverge.com/

June, 3, 2019

YouTube pushes children's videos to pedophiles through content recommendation engine

XENI JARDIN / 8:15 AM MON JUN 3, 2019



https://boingboing.net/2019/06/03/youtube-sexualizes-children.html

YouTube mistakenly links Notre Dame fire to September 11 attacks

April 15, 2019

Even big companies can't always practice what they preach

Artificial Intelligence at Google: Our Principles - Objectives for Al applications

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.
7. Be made available for uses that accord with these principles.

https://ai.google/principles

Next steps

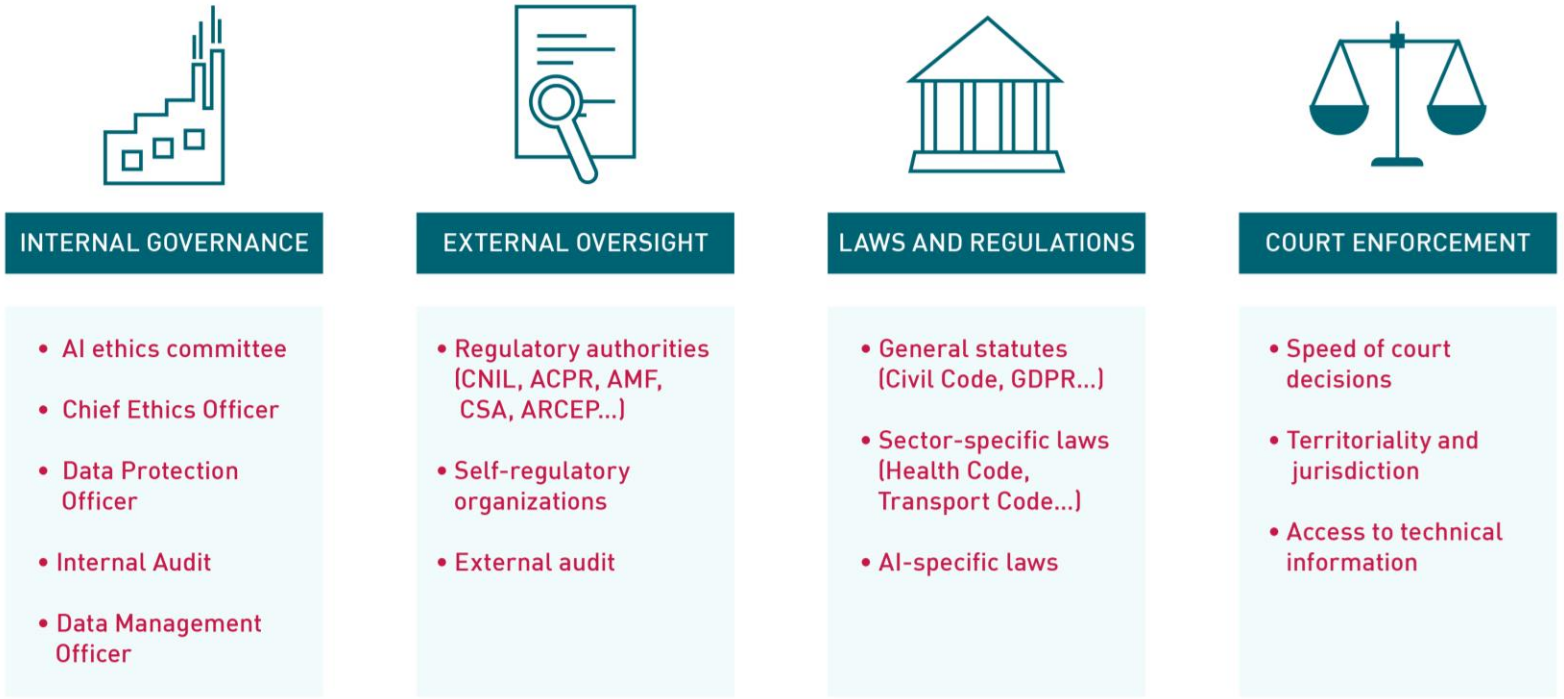
- AI risk management will define the next era of technological advancement and become essential to companies' AI strategies.
- Given AI's rapid development and increasing applications, AI risks are constantly changing and evolving, meaning that comprehensive risk management strategies are needed to avoid reputational damage and facilitate legal compliance.
- **Auditing and testing AI** systems reveal whether issues in a system's development, training, or deployment will lead to biased decision-making. Where any issues are found, these can be addressed with state-of-the-art mitigation techniques. In doing so, organisations can maximise their ability to innovate with confidence.

Auditing AI

- In November 2021, the New York City Council passed a law requiring bias audits for AI tools used by employers in the hiring process.
- Employers and employment agencies based in New York are required to comply with the law starting from July 5, 2023. They need to be aware of their obligations if they use AI to facilitate hiring or promotion decisions.
- For the first time, a city as large as New York will impose fines for undisclosed or biased use of AI, imposing penalties of up to $1,500 per violation on employers and providers.

AI Governance

- **Artificial intelligence governance** is the legal framework for ensuring AI and machine learning technologies are researched and developed with the goal of helping humanity navigate the adoption and use of these systems in ethical and responsible ways.
- AI governance aims to close the gap that exists between accountability and ethics in technological advancement. [Bostrom et al. 2014]

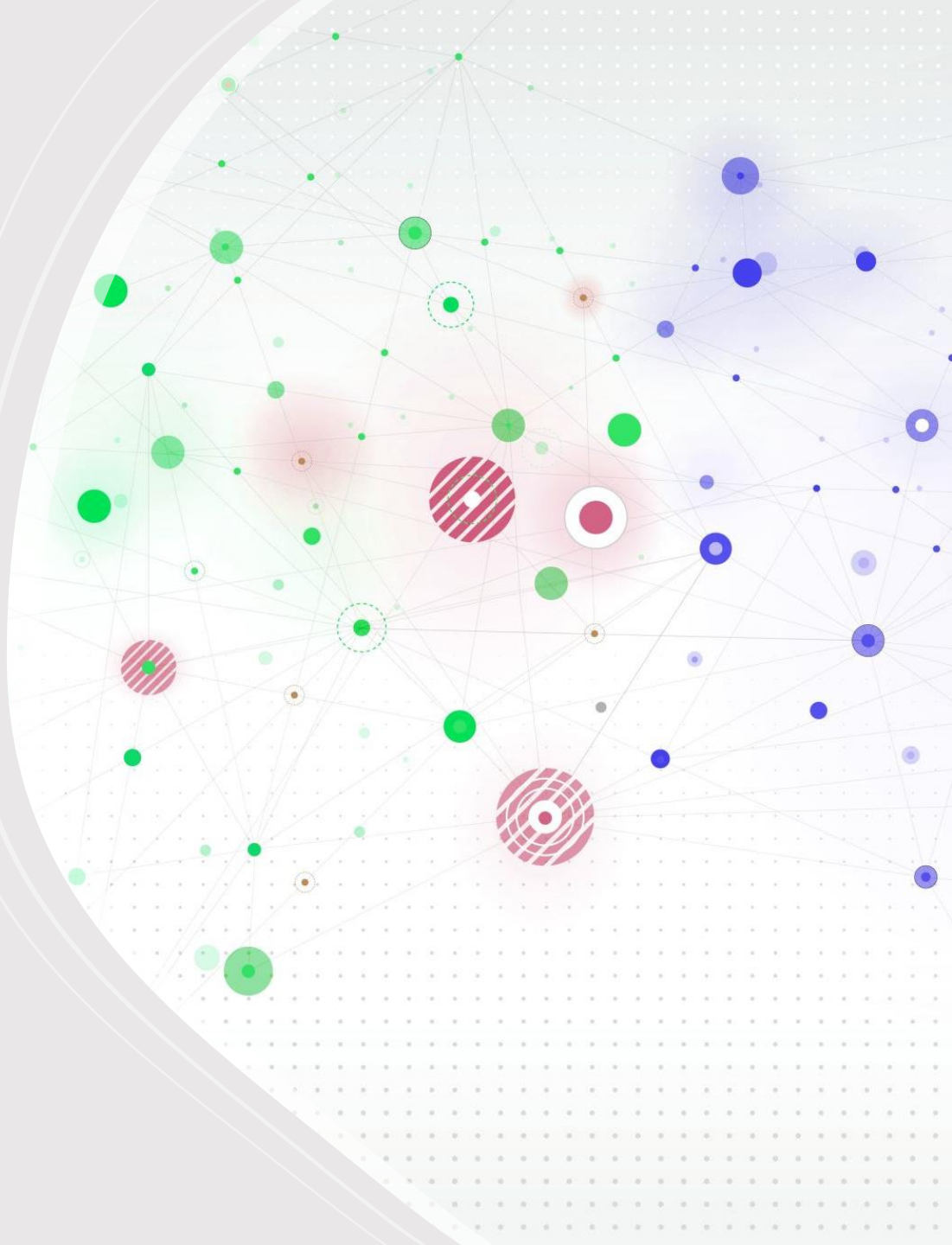


AI governance and regulatory ecosystem

Source: Telecom Paris

4. Impacts of Al on the World of Work

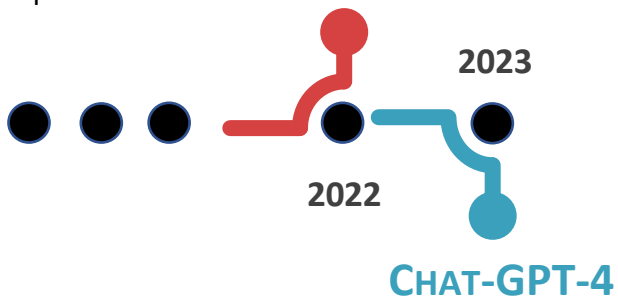
- ▷ 4.1 Task Automation and Job Transformation
- ▷ 4.2 How Artificial Intelligence Will Redefine Management



70 years of History

GENERATIVE AI

- Midjourney: Converting text descriptions into ultra-realistic images.
- Launch of Chat-GPT-3 by OpenAI: A free, standalone version of the conversational agent.
- DALL·E2: DALL·E 2 is an AI system that can create realistic images and art from a description in natural language.
- Stable Diffusion: A technique for generating high-quality images from text descriptions.



Midjourney

ChatGPT

DALL·E 2

CRÉATION DE VISUEL

TEXT PROMPT

strange doodle art by Keith Haring black and white, animal, gustav klimt, totem, alien, egyptian geroglyphic, robot, rainbow, on paper --ar 2:3

Format

AI utilisée : Midjourney

K. Ben Amor – Noway Studio

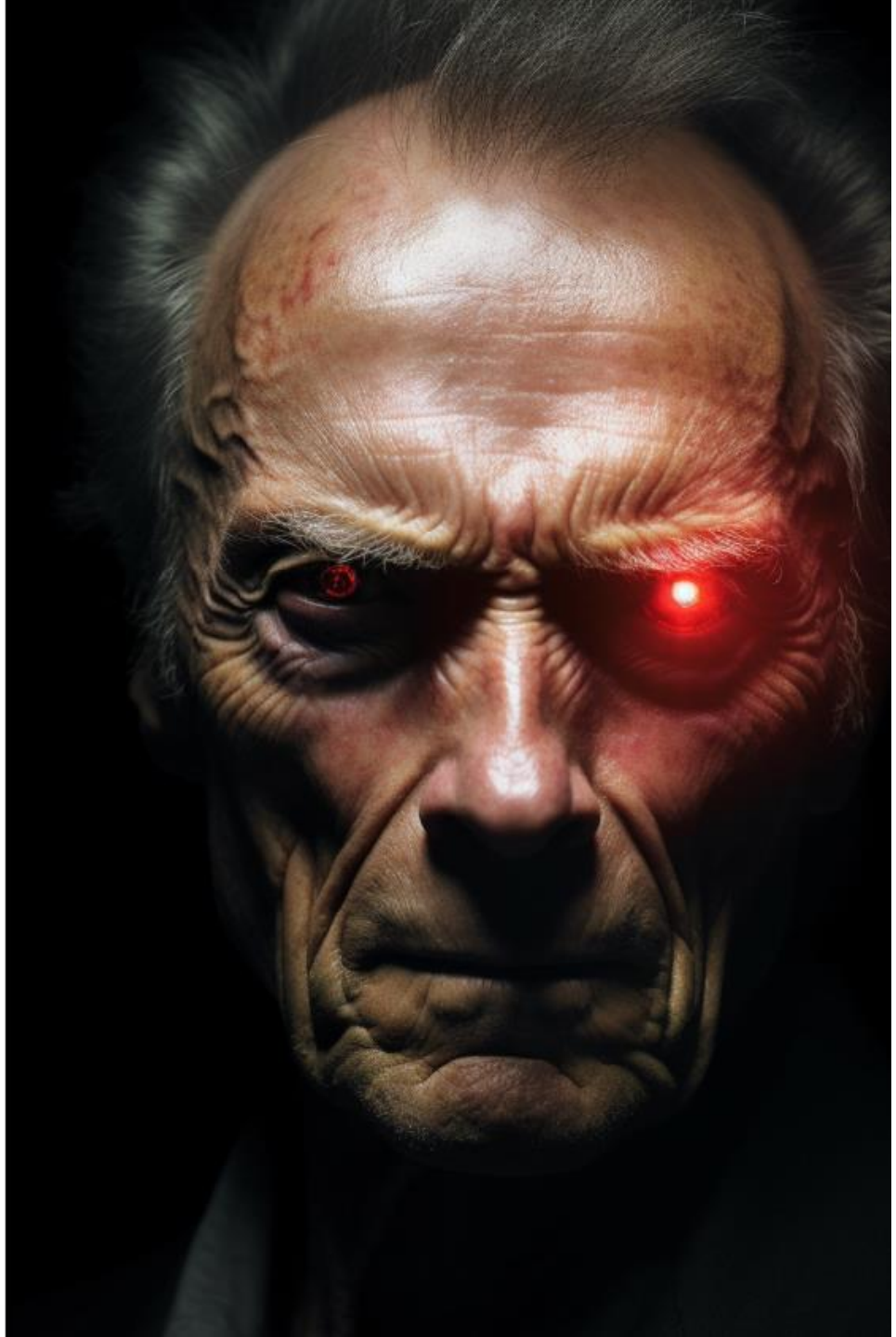
CRÉATION PHOTOGRAPHIQUE

TEXT PROMPT

Clint Eastwood - terminator, with glowing red eye, portrait by Lee Jeffries, real photography, hyperrealistic, 8k --ar 9:16 --v 5.1 --s 50

Format Version Style

AI utilisée : Midjourney



K. Ben Amor – Noway Studio

CRÉATION D'INTÉRIEUR
À PARTIR D'UN CROQUIS

TEXT PROMPT

Raw photo of classic living room, with sphere lighting design, fireplace on the right side, windows on the left side, coated wood parquet, epic living room design, detailed, realistic, natural lighting, beautiful and modern curtains, colored cushions, <lora:epiNoiseoffset_v1:1>

NEGATIVE PROMPT

bad living room, bad interior, bad proportions, blurry, cropped, deformed, error, gross proportions, jpeg artifacts, low quality, lowres, malformed furniture, out of frame, poorly drawn interior, signature, text, ugly, username, watermark, worst quality

AI utilisée : Stable diffusion / Modèle : Realistic Vision / Mode : Text to image

Croquis utilisé

Image obtenue

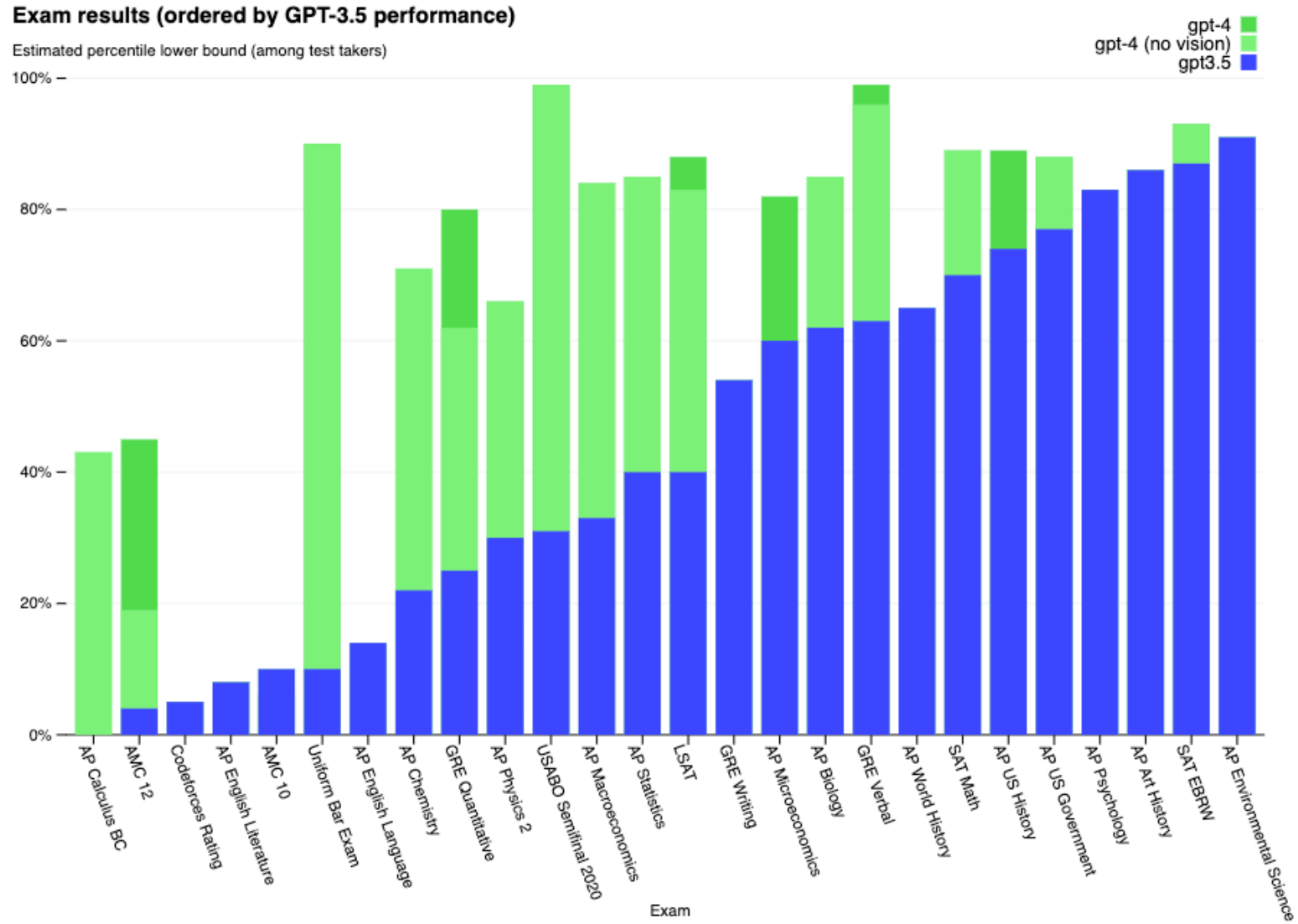
K. Ben Amor – Noway Studio

Boris Eldagsen's AI-generated image titled 'Pseudomnesia: The Electrician' was submitted to the Sony World Photography Awards 2023 and won first prize in the creative open category

- **The artist rejects the award: he accused the Sony World Photograph Awards of failing to distinguish between a photograph and a DALL-E 2-created image, while the organisers condemn a 'deliberate attempt at misleading us'**
- **Many photographers and artists fear their livelihoods are under threat from AI tools that allow anyone to create striking images with just a quick text prompt.**

ChatGPT is going to steal our jobs!

March 2023 : OpenAI announces GPT-4, claims it can beat 90% of humans on the Scholastic Assessment Test (SAT)



Study by OpenAI, OpenResearch, and the University of Pennsylvania on the US Job Market : “GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models “, 2023
https://arxiv.org/pdf/2303.10130.pdf

And what's more, AI resembles us...

AI Avatars: Looking Like Humans, Talking Like Humans, and Moving Like Humans

Virtual Influencer

- Virtual influencer. The computer-generated avatars that take the form of real people. These online personalities are typically created by media agencies and/or brands (Prada, Samsung, Balmain etc.).
- Recent reports have shown that virtual influencers have nearly 3x higher engagement rates than ‘real’ influencers

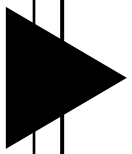
LilMiquela is one of the most well-known virtual influencers.

Stages of AI

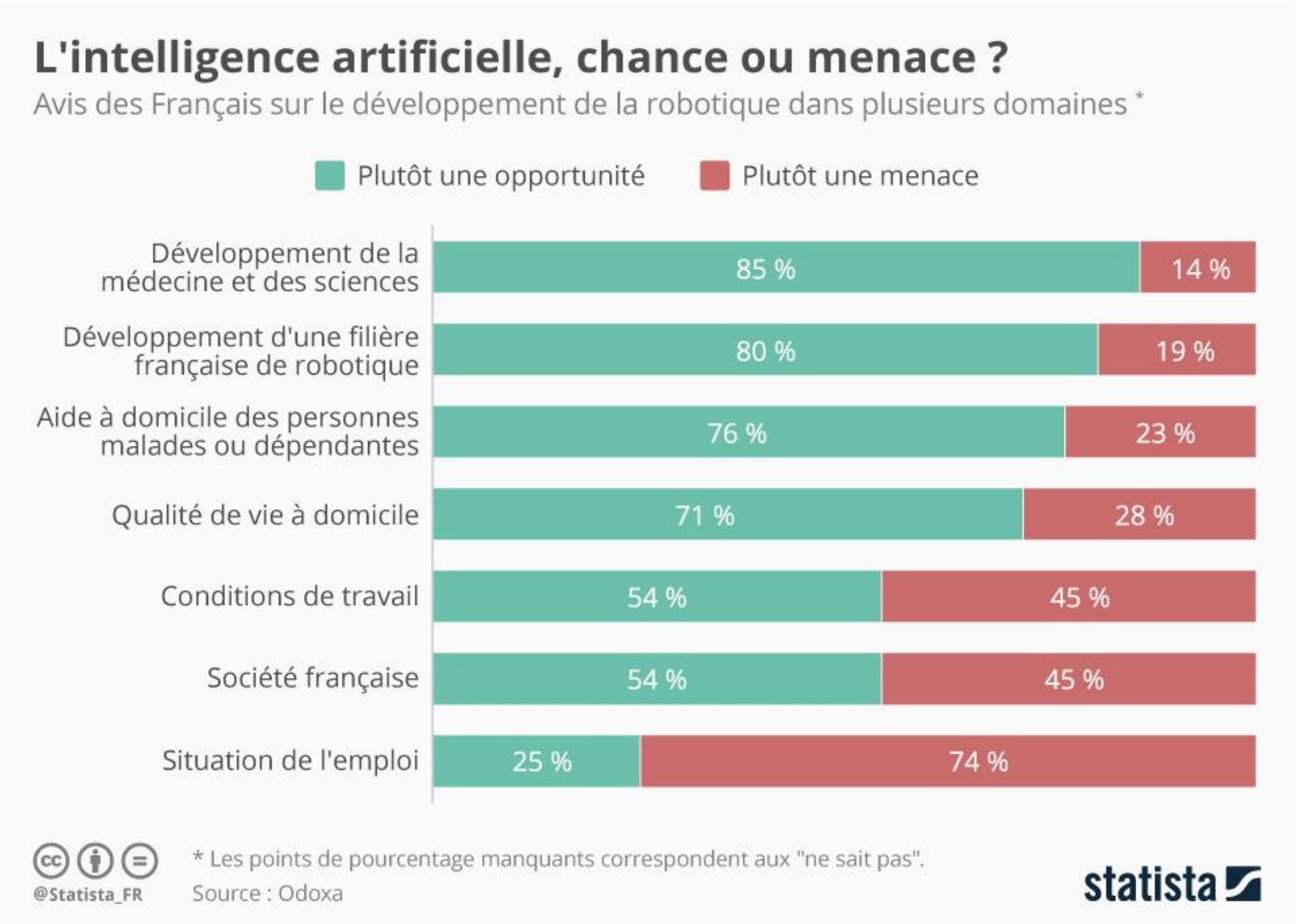
NARROW AI

WEAK AI

Specialized in specific and limited tasks



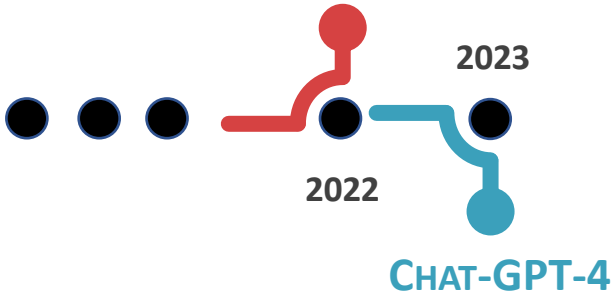
AI opportunity or threat ?



70 years of History

GENERATIVE AI

- Midjourney: Converting text descriptions into ultra-realistic images.
- Launch of Chat-GPT-3 by OpenAI: A free, standalone version of the conversational agent.
- DALL·E2: DALL·E 2 is an AI system that can create realistic images and art from a description in natural language.
- Stable Diffusion: A technique for generating high-quality images from text descriptions.



- March 2023: Elon Musk and hundreds of experts call for a pause in AI, citing "major risks to humanity."
- May 2023: Geoffrey Hinton resigns from Google to sound the alarm about AI.

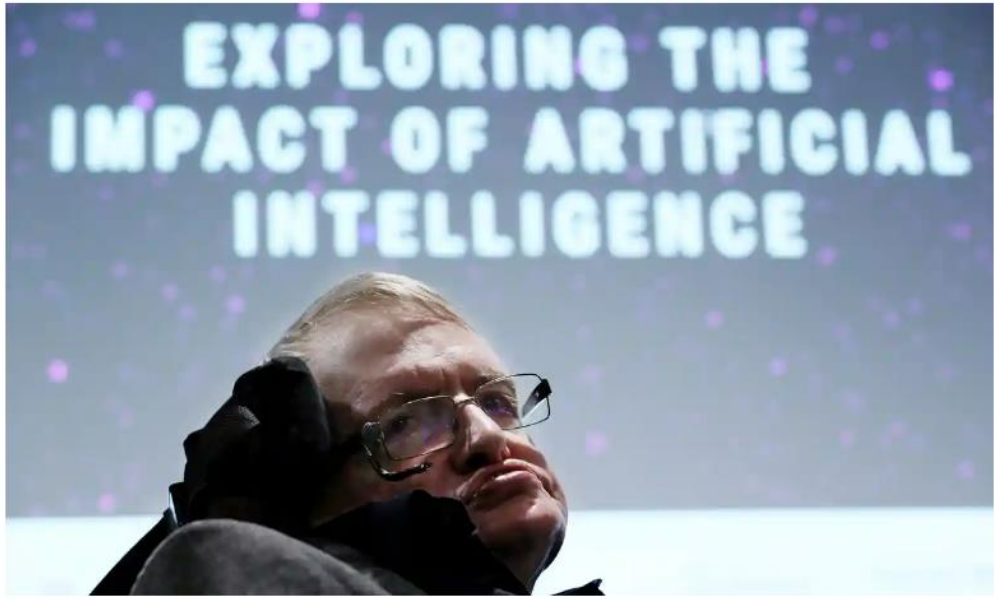
Are we doomed ?

https://www.theguardian.com/ - oct 2016

This article is more than **6 years old**

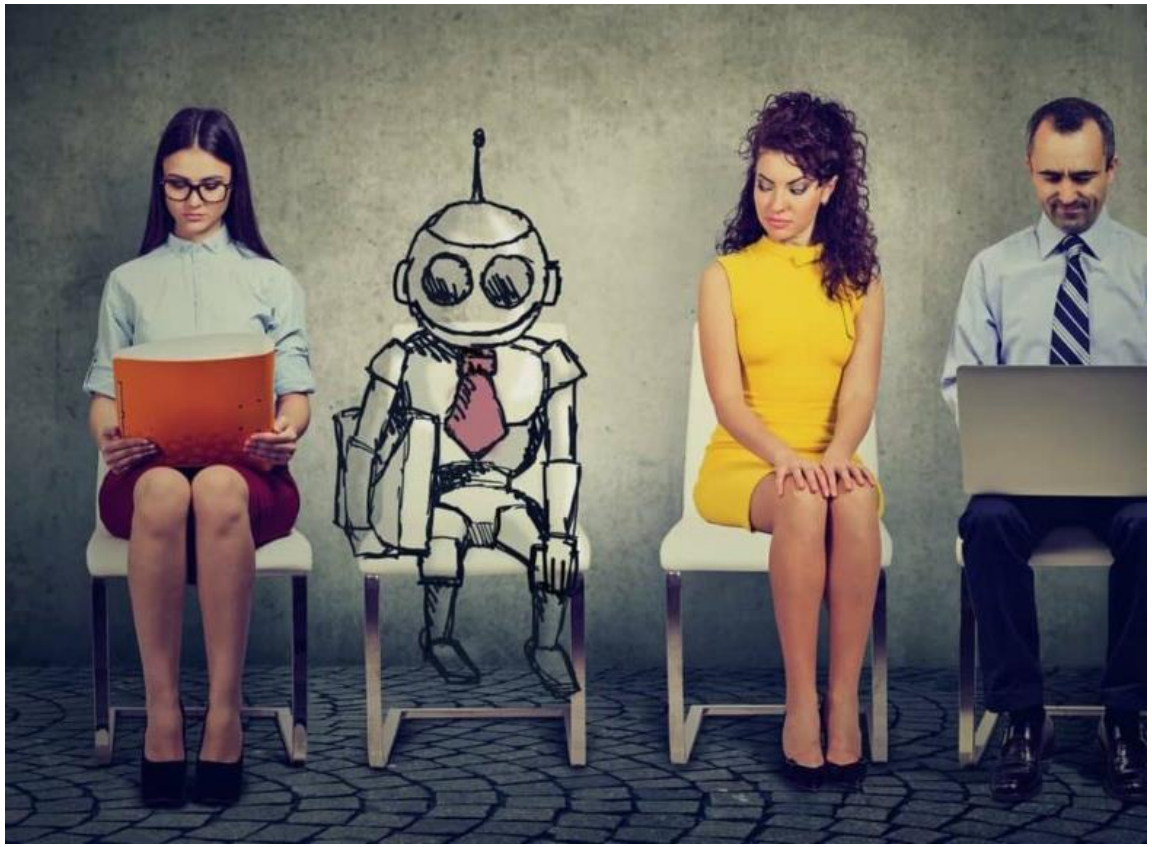
Stephen Hawking: AI will be 'either best or worst thing' for humanity

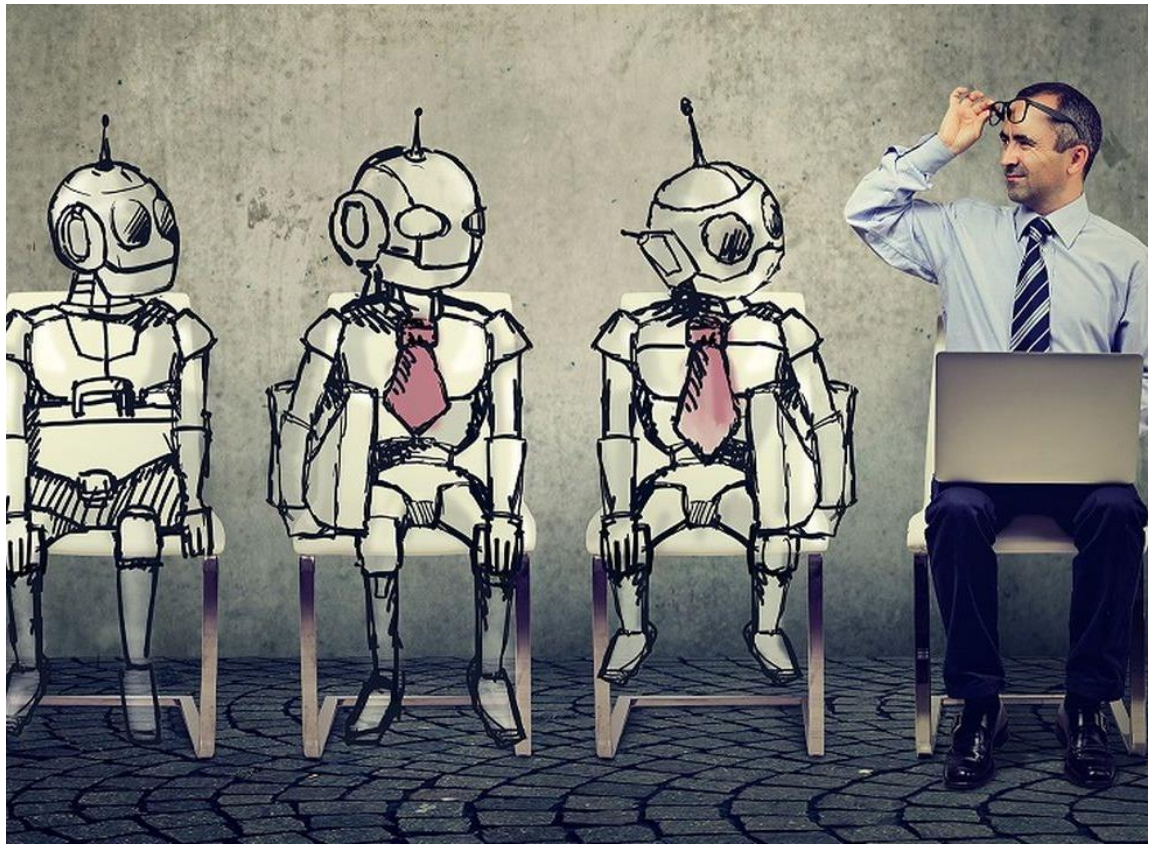
Professor praises creation of Cambridge University institute to study future of artificial intelligence



Stephen Hawking at the opening of the Leverhulme Centre for the Future of Intelligence on Wednesday. Photograph: Chris Radburn/PA

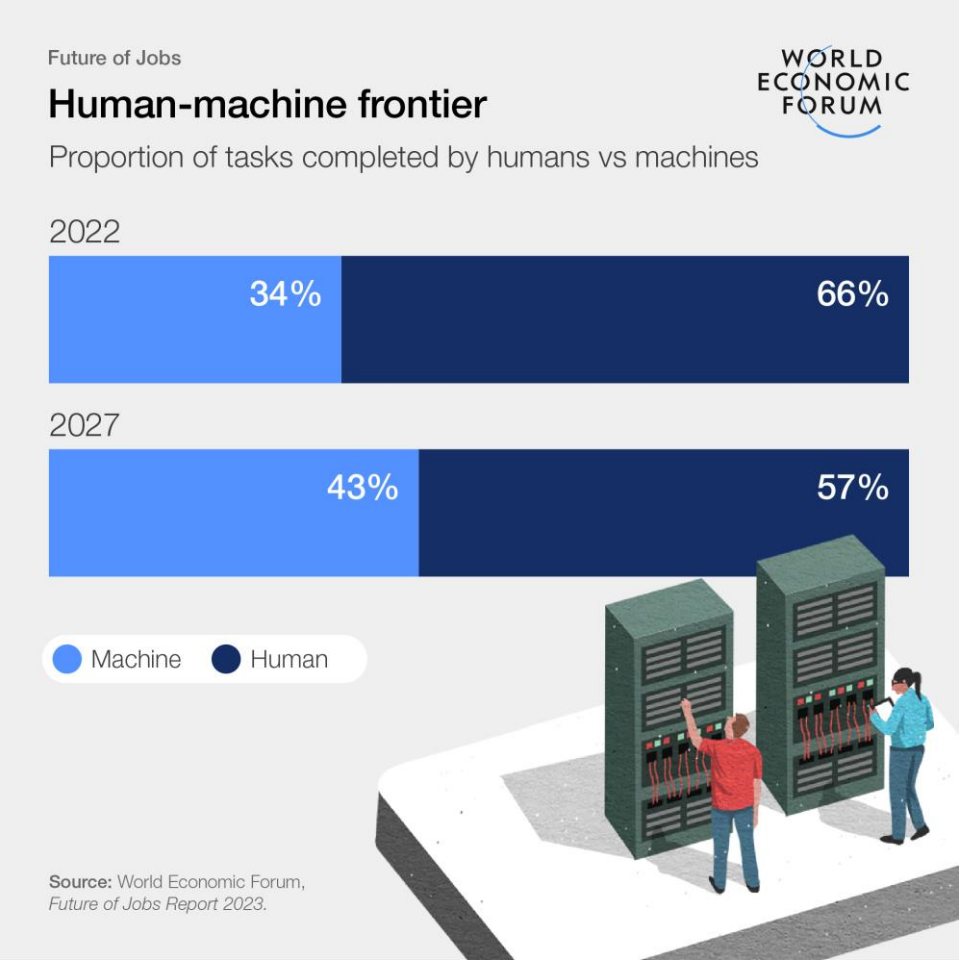
Is our job at risk?





Futur of jobs 2023

World Economic Forum (WEF) - May 2023



Forbes

FORBES > LEADERSHIP > CAREERS

EDITORS' PICK

Goldman Sachs Predicts 300 Million Jobs Will Be Lost Or Degraded By Artificial Intelligence

Jack Kelly Senior Contributor

I write actionable interview, career and salary advice.

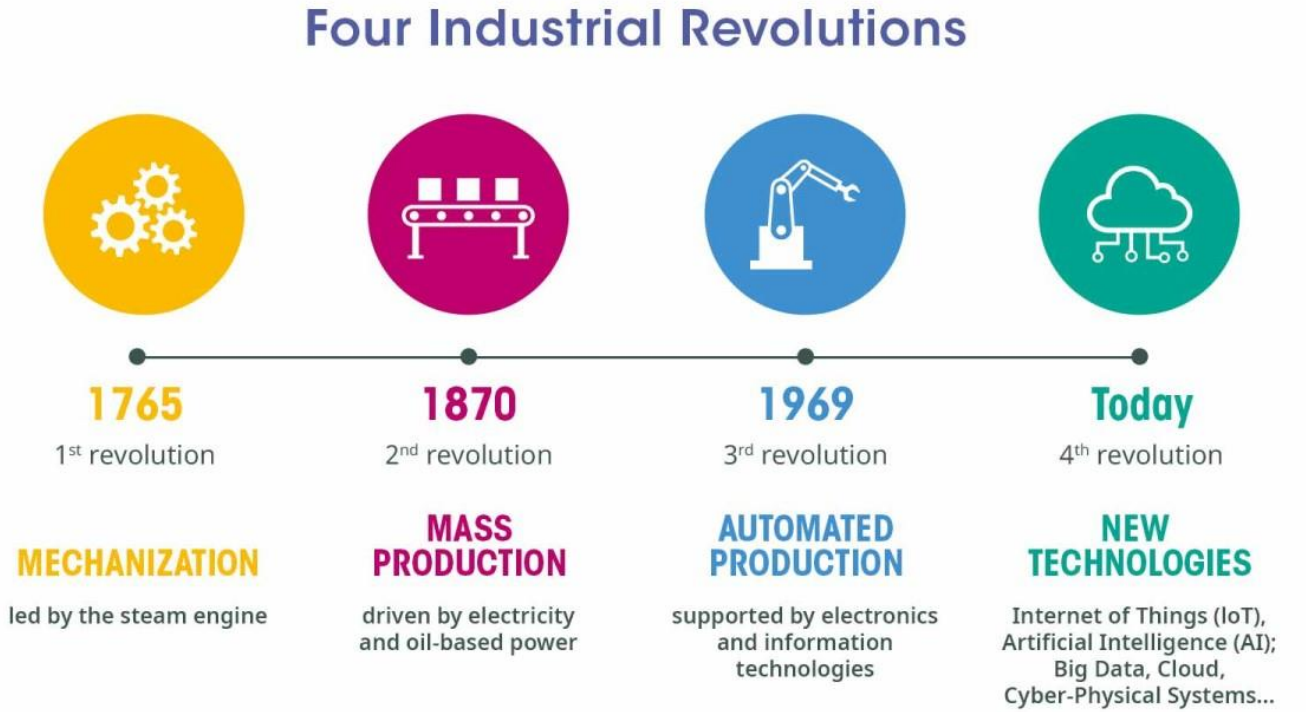
Follow

4

Mar 31, 2023, 10:48am EDT

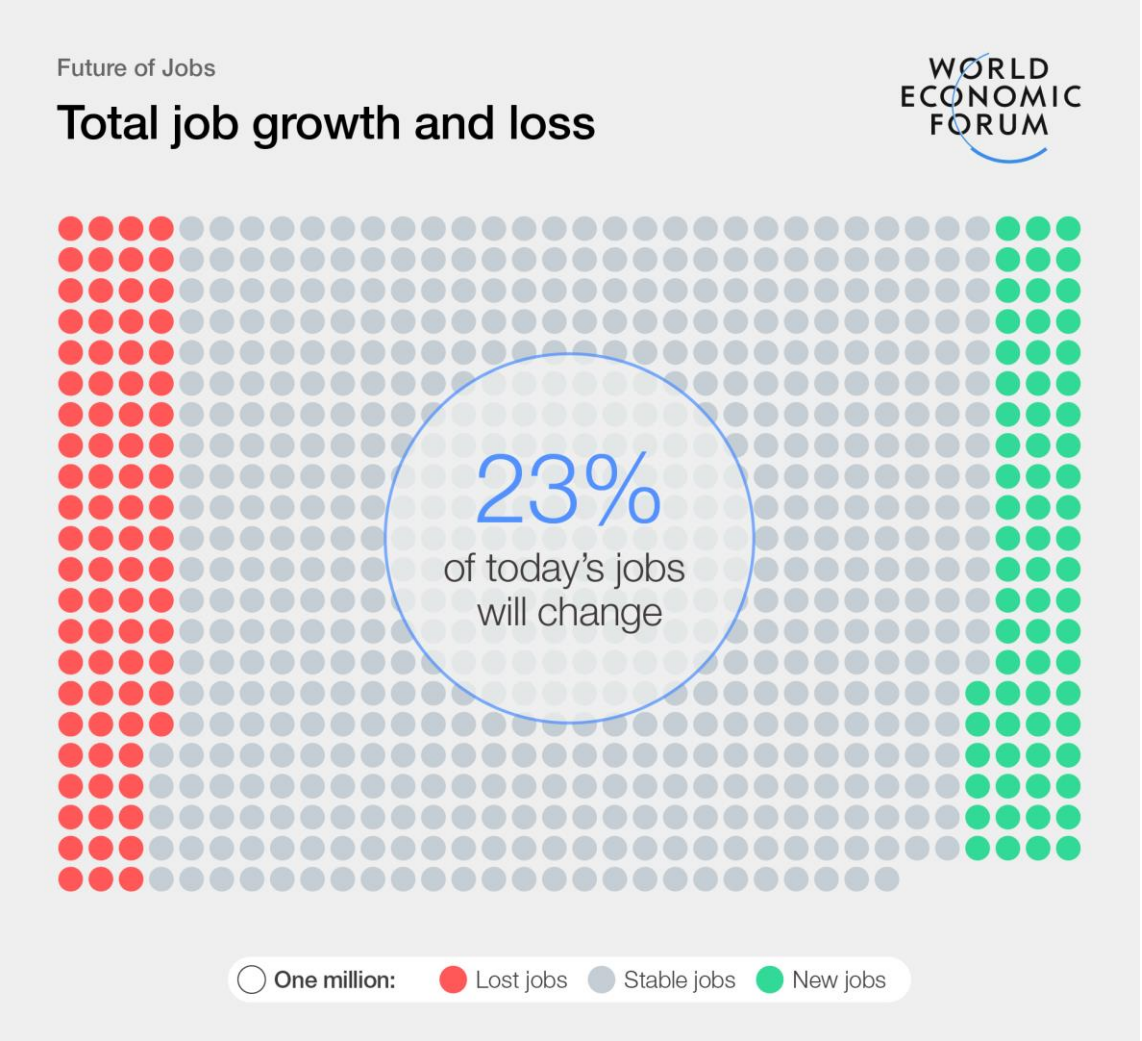
AI will not destroy more jobs than it will create

- History has time and again demonstrated that the loss of jobs resulting from technological advancements is often counterbalanced by the creation of new jobs.
- Throughout different eras, from the Industrial Revolution to the digital age, disruptive technologies have indeed displaced certain occupations. However, they have also given rise to entirely new industries, transforming the job market and creating a demand for new skills and expertise.



AI will not destroy more jobs than it will create

- 60% of workers today are employed in occupations that did not exist in the 1940s, which means that over 85% of employment growth over the past 80 years is attributed to the technology-driven creation of new positions.
- Similarly, such would also be the case with AI's emergence.
- According to the World Economic Forum, AI will be able to create at least 12 million more jobs more than it destroys by 2025.



Fastest growing vs. fastest declining jobs

Top 10 fastest growing jobs

1.	AI and Machine Learning Specialists
2.	Sustainability Specialists
3.	Business Intelligence Analysts
4.	Information Security Analysts
5.	Fintech Engineers
6.	Data Analysts and Scientists
7.	Robotics Engineers
8.	Electrotechnology Engineers
9.	Agricultural Equipment Operators
10.	Digital Transformation Specialists

Top 10 fastest declining jobs

1.	Bank Tellers and Related Clerks
2.	Postal Service Clerks
3.	Cashiers and ticket Clerks
4.	Data Entry Clerks
5.	Administrative and Executive Secretaries
6.	Material-Recording and Stock-Keeping Clerks
7.	Accounting, Bookkeeping and Payroll Clerks
8.	Legislators and Officials
9.	Statistical, Finance and Insurance Clerks
10.	Door-To-Door Sales Workers, News and Street Vendors, and Related Workers

Source
World Economic Forum, Future of Jobs Report 2023.

Note
The jobs which survey respondents expect to grow most quickly from 2023 to 2027 as a fraction of present employment figures

Faster growing jobs

- The development, implementation, and maintenance of AI systems by itself would require a skilled workforce specialising in areas such as **AI and ML specialists, data analysis, information security**.
- Robots will inevitably suffer glitches, need updates and require new parts. As companies rely more and more on automation, they will require more people with technical skills to maintain, replace, update and fix AI technology.
- AI can potentially generate new jobs by enabling new sectors and business models. For instance in Singapore, new roles including AI prompt engineers have cropped up. These prompt engineers create and curate prompts for chatbots like ChatGPT and can earn an upwards of US$335,000 (S$445,000) per annum.

Declining jobs

- Some of the jobs that are most at risk of being replaced by AI-powered systems **include routine and repetitive roles**, such as those involving manufacturing, administration and customer service.
- By 2030, up to 20 million manufacturing jobs globally will be lost to robots by 2030, while up to 46 per cent of office and administrative support jobs could also possibly be automated.
- AI's capabilities are not just limited to automating routine tasks (red): for example for Data Entry Clerks, AI-powered tools can quickly scan and extract data from documents
- Advancements in NLP and ML algorithms have enabled AI systems to perform complex cognitive tasks, including data analysis, decision-making, and even creative endeavours : As a result, professions that were once considered secure, such as accounting, legal services, and even some aspects of medical diagnostics, are now at risk of being taken over by AI. For example for Accountants and Auditors, AI-powered tools can analyse financial data and generate reports

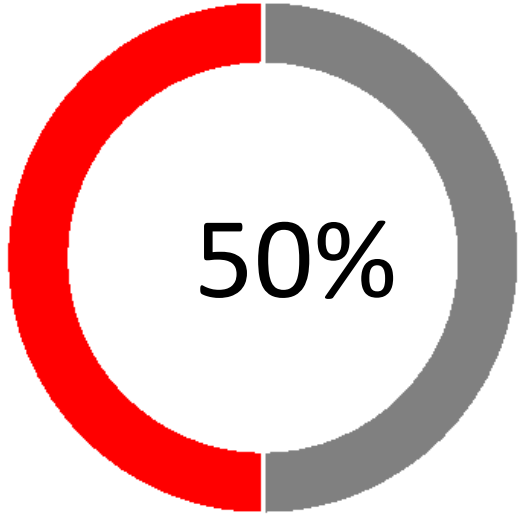
How Artificial Intelligence Will Redefine Management ?

- November 2016: Researchers from Accenture conducted a survey involving 1,770 managers from 14 countries and interviewed 37 executives responsible for digital transformation in their organizations (*).
- Based on this data, they identified five practices that successful managers will need to master:
  - Practice 1: Letting AI Handle Administration
  - Practice 2: Focus on Judgment Work
  - Practice 3: Treating Intelligent Machines as "Colleagues"
  - Practice 4: Working as a Designer
  - Practice 5: Developing Social Skills and Networks
- We correlated these practices with a recent study on Intelligent Automation - Learn How to Harness Artificial Intelligence to Boost Business & Make Our World More Human [Bornet et al. 2020] and the results on the Future of Jobs 2023 by the World Economic Forum (WEF) - May 2023.

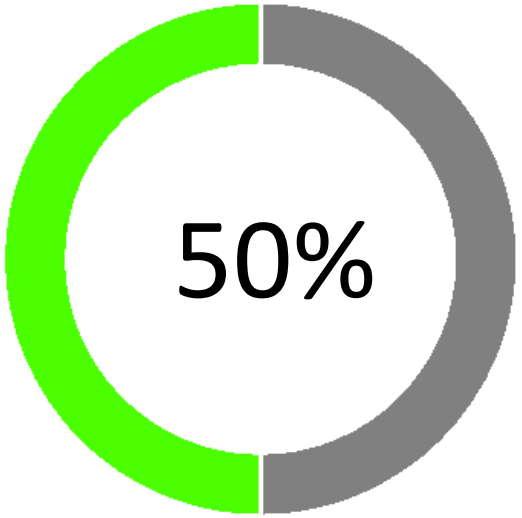
(*) https://hbr.org/2016/11/how-artificial-intelligence-will-redefine-management

Practice 1: Leave Administration to AI

Average percentage of time spent by employees	% Time Current
Individual thinking, Research, Administrative Time	42%
Administrative Tasks	16%
Productive thinking Time	10%
Productive time doing work – non routine	8%
Productive time doing work –routine	8%
Collaboration – internal & external	49%
Unproductive messages – reading /writing	14%
Unproductive meetings	12%
Productive collaboration (e.g. meetings, workshops)	9%
Productive emails	9%
Socialization	5%
Others / non-work / social media / interruptions	9%
Total	100%



high proportion of time spent on unproductive, routine, and administrative activities (highlighted in red)



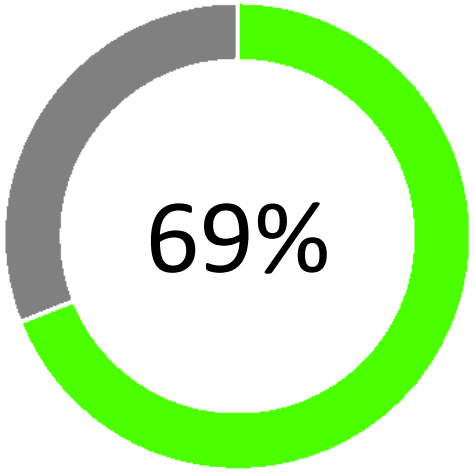
more than 50% of employees, wasteful meetings and excessive emails are reported to be the main obstacle in their work

Source: "Intelligent Automation - Learn How to Harness Artificial Intelligence to Boost Business & Make Our World More Human" Date de publication : Octobre 2020 Pascal Bornet, Jochen Wirtz

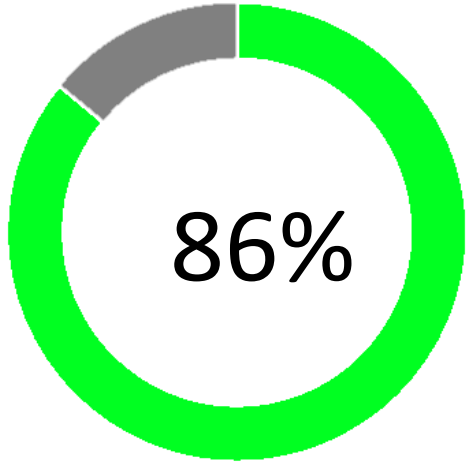
Practice 1: Leave Administration to AI

Average percentage of time spent by employees	% Time Current	% Time Ideal
Individual thinking, Research, Administrative Time	42%	45%
Administrative Tasks	16%	
Productive thinking Time	10%	25%
Productive time doing work – non routine	8%	20%
Productive time doing work –routine	8%	
Collaboration – internal & external	49%	55%
Unproductive messages – reading /writing	14%	
Unproductive meetings	12%	
Productive collaboration (e.g. meetings, workshops)	9%	25%
Productive emails	9%	20%
Socialization	5%	10%
Others / non-work / social media / interruptions	9%	
Total	100%	100%

The question is then, besides removing tedious tasks, what would make people happier and more engaged at work?



of employees expect that automation will give them more time to do their primary job duties



of employees think the use of automation in the workplace will let them think of work in new and innovative ways

Practice 1: Leave Administration to AI

Average percentage of time spent by employees	% Time Current	% Time Ideal
Individual thinking, Research, Administrative Time	42%	45%
Administrative Tasks	16%	
Productive thinking Time	10%	25%
Productive time doing work – non routine	8%	20%
Productive time doing work –routine	8%	
Collaboration – internal & external	49%	55%
Unproductive messages – reading /writing	14%	
Unproductive meetings	12%	
Productive collaboration (e.g. meetings, workshops)	9%	25%
Productive emails	9%	20%
Socialization	5%	10%
Others / non-work / social media / interruptions	9%	
Total	100%	100%

- Studies have found that people like to have opportunities to solve problems in their work.
- Also, variety among tasks at work leads to increased happiness and higher productivity.
- Employee experience would be improved by switching to the “ideal” daily division of activities.

This ideal time division would also fit well with the imperative of a profitable company, as employees would focus on the work activities which generate the most value for companies (i.e., productive tasks).

Potential levers to improve employee experience and achieve the ideal distribution:

			Opportunity for transformation		
Average percentage of time spent by employees	% Time Current	% Time Ideal	Automate	Augment	Abandon
Individual thinking, Research, Administrative Time	42%	45%			
Administrative Tasks	165%				
Productive thinking Time	10%	25%			
Productive time doing work – non routine	8%	20%			
Productive time doing work –routine	8%				
Collaboration – internal & external	49%	55%			
Unproductive messages – reading /writing	14%				
Unproductive meetings	12%				
Productive collaboration (e.g. meetings, workshops)	9%	25%			
Productive emails	9%	20%			
Socialization	5%	10%			
Others / non-work / social media / interruptions	9%				
Total	100%	100%			

Automate: companies should identify and automate routine activities, such as generating a PowerPoint presentation for a weekly meeting or recording invoices in accounting software

Potential levers to improve employee experience and achieve the ideal distribution:

			Opportunity for transformation		
Average percentage of time spent by employees	% Time Current	% Time Ideal	Automate	Augment	Abandon
Individual thinking, Research, Administrative Time	42%	45%			
Administrative Tasks	165%				
Productive thinking Time	10%	25%			
Productive time doing work – non routine	8%	20%			
Productive time doing work –routine	8%				
Collaboration – internal & external	49%	55%			
Unproductive messages – reading /writing	14%				
Unproductive meetings	12%				
Productive collaboration (e.g. meetings, workshops)	9%	25%			
Productive emails	9%	20%			
Socialization	5%	10%			
Others / non-work / social media / interruptions	9%				
Total	100%	100%			

Augment: organizations should seize the opportunity to increase the value of work activities delivered by employees. IA is used as a crucial component here, with, for example, the generation of insights through advanced analytics to help decision making

Potential levers to improve employee experience and achieve the ideal distribution:

			Opportunity for transformation		
Average percentage of time spent by employees	% Time Current	% Time Ideal	Automate	Augment	Abandon
Individual thinking, Research, Administrative Time	42%	45%			
Administrative Tasks	165%				
Productive thinking Time	10%	25%			
Productive time doing work – non routine	8%	20%			
Productive time doing work –routine	8%				
Collaboration – internal & external	49%	55%			
Unproductive messages – reading /writing	14%				
Unproductive meetings	12%				
Productive collaboration (e.g. meetings, workshops)	9%	25%			
Productive emails	9%	20%			
Socialization	5%	10%			
Others / non-work / social media / interruptions	9%				
Total	100%	100%			

Abandon: some work activities do not fit with leading practices for efficient work, and represent an obstacle to the employee's experience. These activities should be reduced or eliminated. For example, restricting the volume of meetings and email traffic is essential

« Innovate more to produce more... rather than working more to produce more»

			Opportunity for transformation		
Average percentage of time spent by employees	% Time Current	% Time Ideal	Automate	Augment	Abandon
Individual thinking, Research, Administrative Time	42%	45%	22%	15%	5%
Administrative Tasks	16%		11%		5%
Productive thinking Time	10%	25%	3%	7%	
Productive time doing work – non routine	8%	20%		8%	
Productive time doing work –routine	8%		8%		
Collaboration – internal & external	49%	55%	17%	17%	15%
Unproductive messages – reading /writing	14%		6%		8%
Unproductive meetings	12%		5%		7%
Productive collaboration (e.g. meetings, workshops)	9%	25%	2%	7%	
Productive emails	9%	20%	3%	6%	
Socialization	5%	10%	1%	4%	
Others / non-work / social media / interruptions	9%		3%		6%
Total	100%	100%	**42%**	**32%**	**26%**

Some useful tools

Practice 2: Focus on Judgment Work

- Many decisions require insight beyond what artificial intelligence can squeeze from data alone.
- Managers use their knowledge of organizational history and culture, as well as empathy and ethical reflection.

This is the essence of human judgment — the application of experience and expertise to critical business decisions and practices.

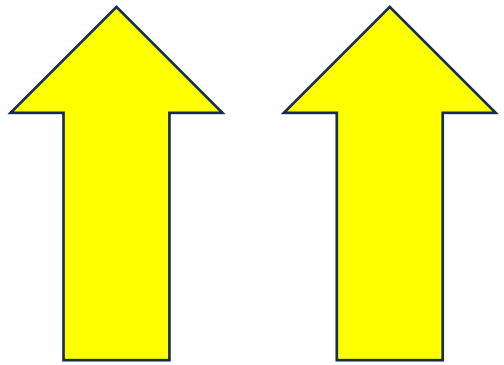
Average percentage of time spent by employees	% Time Current	% Time Ideal
Individual thinking, Research, Administrative Time	42%	45%
Administrative Tasks	16%	
Productive thinking Time	10%	25%
Productive time doing work – non routine	8%	20%
Productive time doing work –routine	8%	
Collaboration – internal & external	49%	55%
Unproductive messages – reading /writing	14%	
Unproductive meetings	12%	
Productive collaboration (e.g. meetings, workshops)	9%	25%
Productive emails	9%	20%
Socialization	5%	10%
Others / non-work / social media / interruptions	9%	
Total	100%	100%

Practice 2: Focus on Judgment Work

Among the skills that organizations consider essential to their work, cognitive skills are at the top

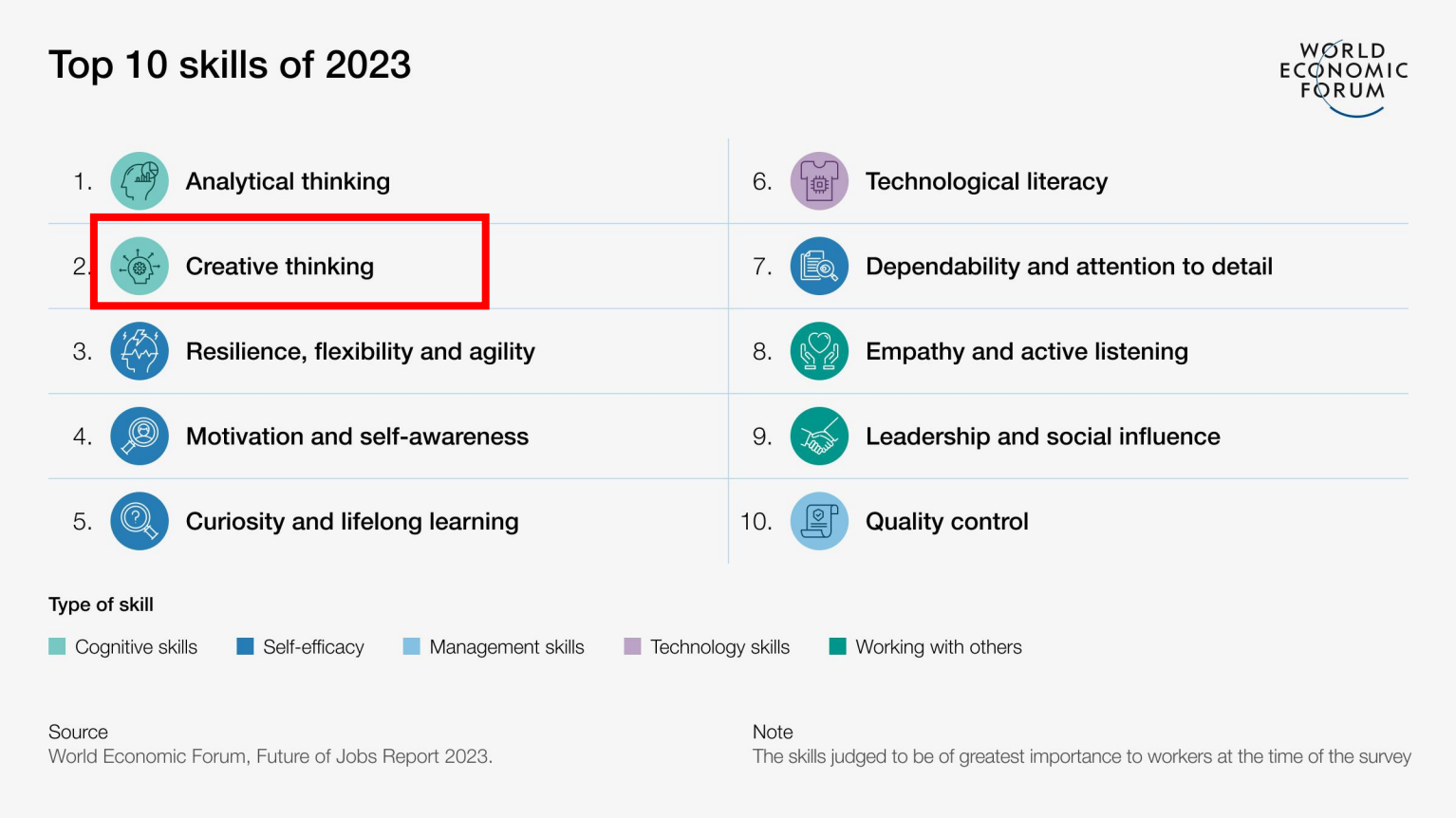
Practice 3 - Treat Intelligent Machines as « Colleagues »

		Opportunity for transformation		
Average percentage of time spent by employees	% Time Current	Automate	Augment	Abandon
Individual thinking, Research, Administrative Time	42%	22%	15%	5%
Collaboration – internal & external	49%	17%	17%	15%
Others / non-work / social media / interruptions	9%	3%		6%
Total	100%	**42%**	**32%**	**26%**



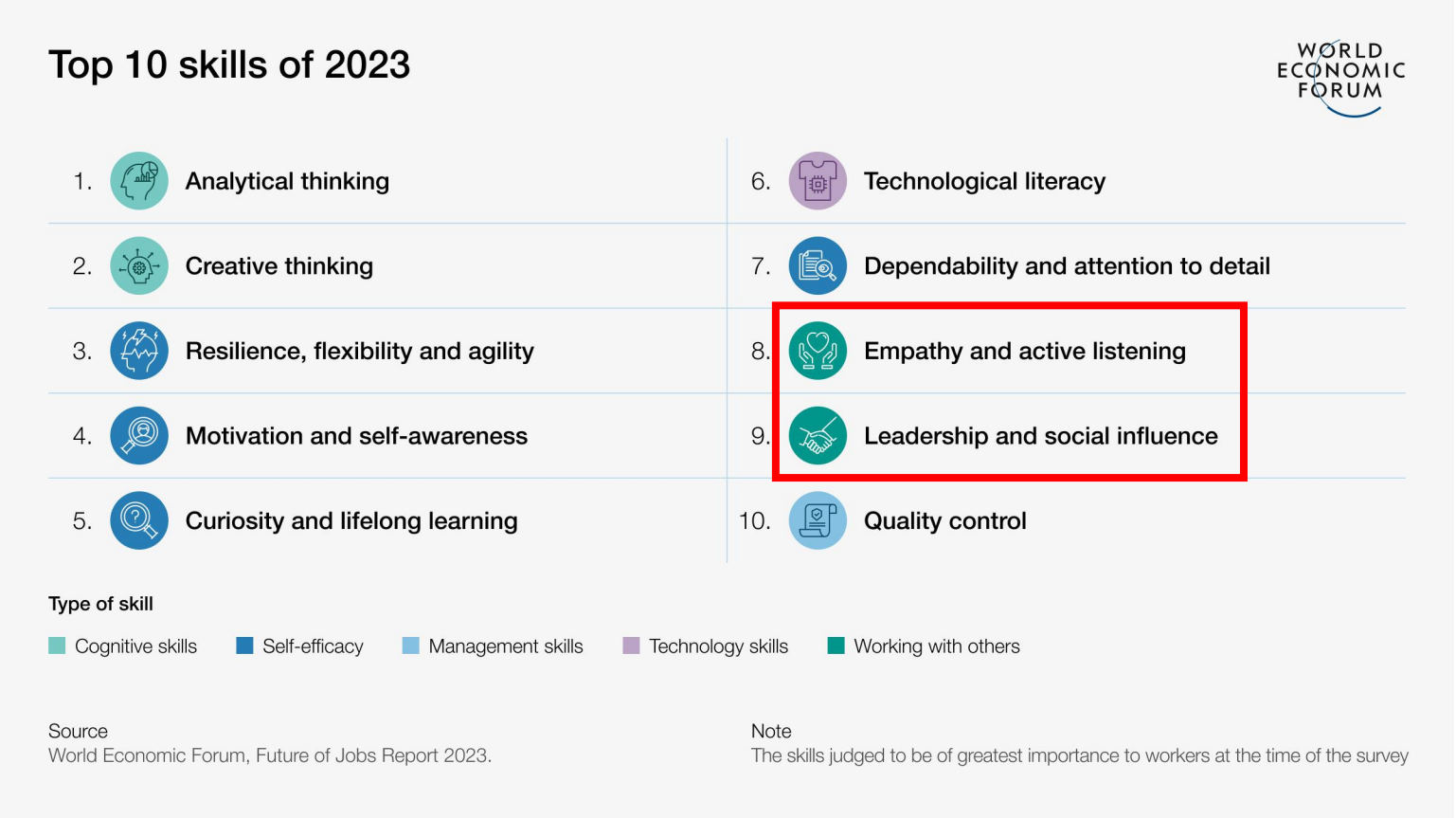
- Highlight the collaboration potential between managers and AI technologies.
- Mention intelligent machines' role in decision support, data-driven simulations, and search activities

Practice 4 - Work Like a Designer



Creative thinking and experimentation are key skill area managers need to learn to stay successful as AI increasingly takes over administrative work.

Practice 5 - Develop Social Skills and Networks

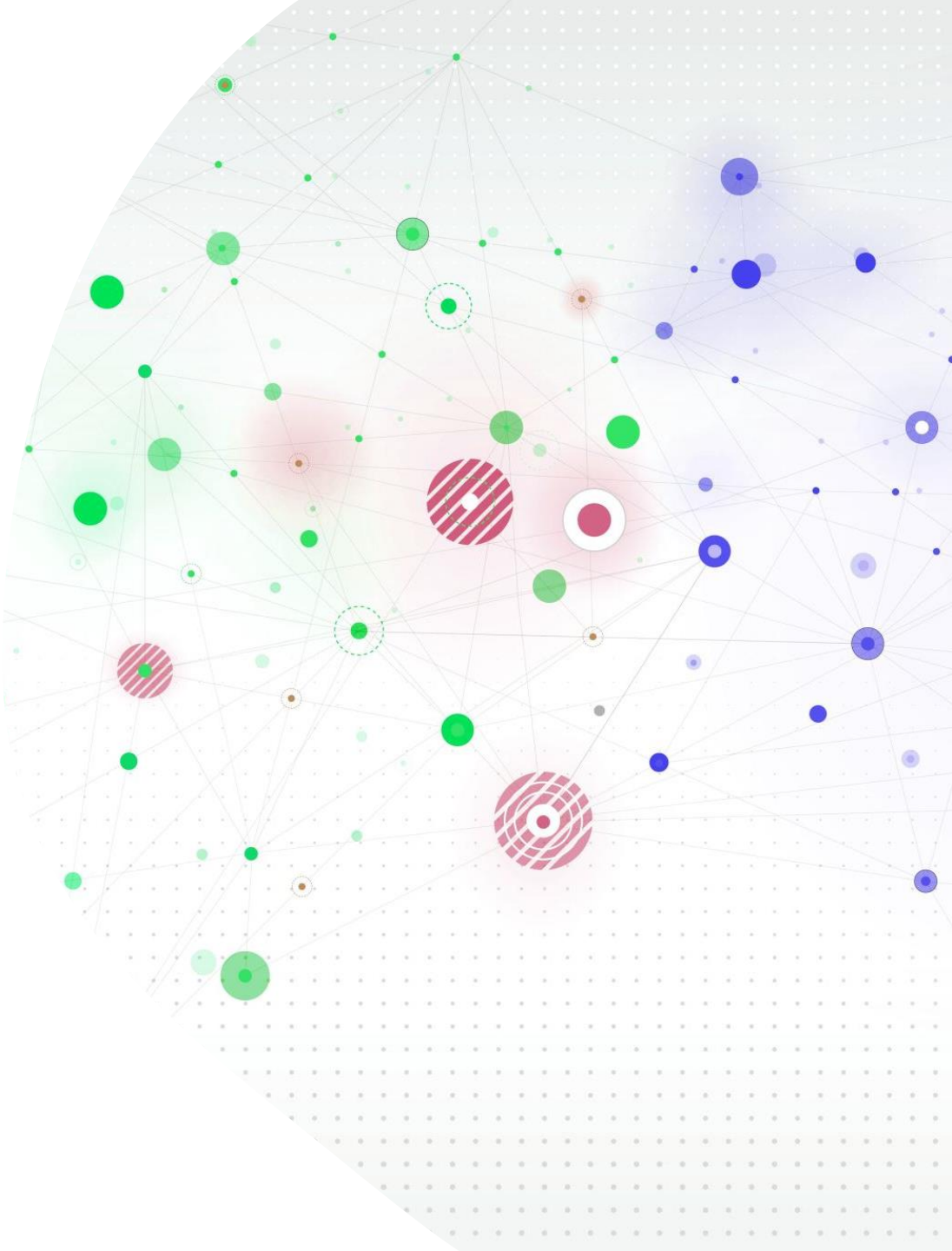


- In order to use digital technologies to tap into the knowledge and judgment of their partners, customers, and communities, managers must be able to discover and bring together diverse perspectives, ideas, and experiences
- This envolves working with others skills

5. Managing Artificial Intelligence

▷Monitoring and improvement of Al system performance

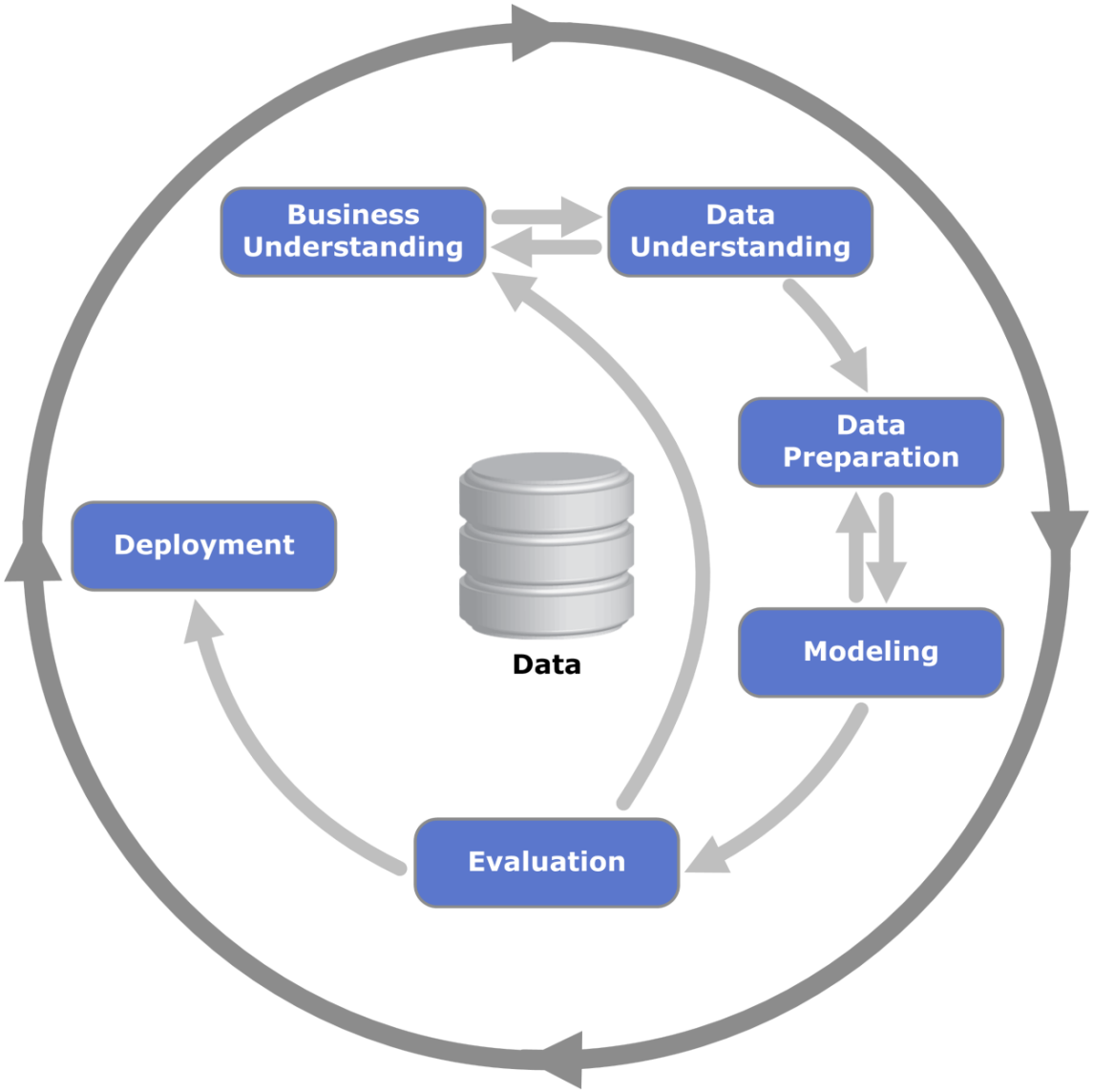
▷ Future Frontiers in Managing Al



Managing AI

Traditional data science activities —are well captured with the CRISP-DM methodology — include executing the following steps in an iterative manner until project goals are achieved:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Model Training (or Modeling)
5. Model Evaluation
6. Model Deployment



Improving enterprise data science throughput requires expanding the AI Model Lifecycle Management beyond the CRISP-DM steps.

This expanded framework incorporates additional tasks that are essential for effectively managing AI projects. In particular, it includes activities related to data collection, data governance on the frontend (before training AI models), and the monitoring of deployed AI models for quality, fairness, and explainability on the backend (after AI models are deployed).

Managing AI

Before CRISP-DM:

1. **Business Opportunity Identification**:
   - Assessing business needs and problems to solve.
   - Determining expected benefits and available resources.
2. **AI Strategy Development:**
   - Risk assessment and priority definition.

CRISP-DM Steps

After CRISP-DM:

1. **Monitoring Deployed AI Models:**
   - Establishing monitoring mechanisms to track model performance.
   - Continuously assessing model quality, fairness, and explainability.
   - Detecting concept drift and addressing issues that may arise.
2. **Model Maintenance and Improvement:**
   - Periodically updating models with new data and addressing changing business needs.
   - Optimizing model performance and addressing biases or fairness concerns.
   - Enhancing model explainability to improve transparency and stakeholder trust.
3. **Model Retirement:**
   - Planning and removing obsolete AI models.
   - Ensuring continuity of services provided by retired models.
4. **Continuous Improvement:**
   - Utilizing acquired knowledge to enhance processes and future outcomes.
   - Exploring new opportunities and innovations in AI.

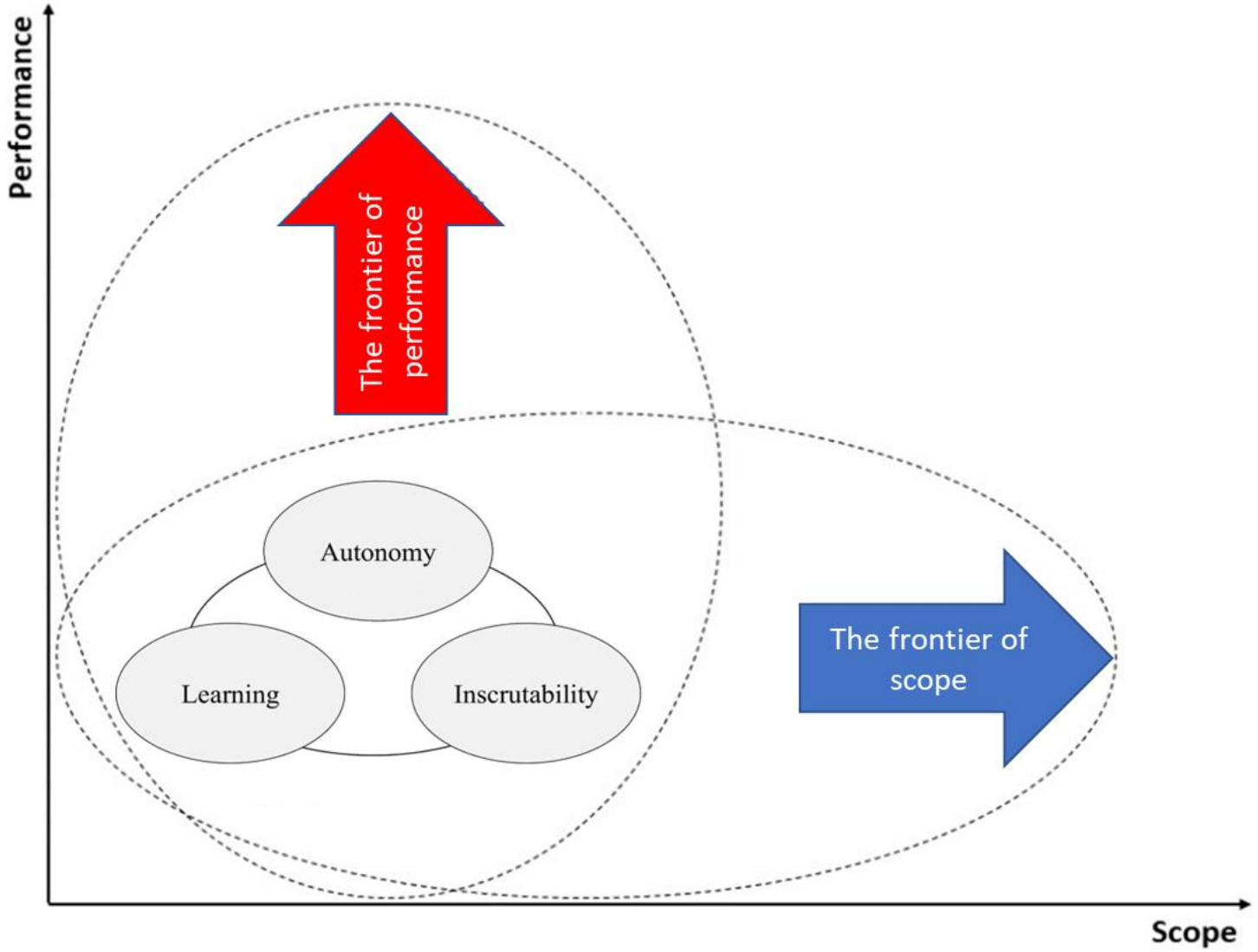
Managing AI

- Note that the steps involved in managing AI projects can vary based on the specific context of each project. It is important to adapt these steps according to the unique needs and objectives of your AI project.
- For example, data scientists can leverage low-code environments to develop pipelines for data preprocessing, feature engineering, and model development.
- Additionally, the emergence of AutoAI/AutoML tools automates various aspects of the AI pipeline, including feature transformation, feature engineering, algorithm selection, and hyperparameter optimization.
- Adaptation and automation are key factors in improving efficiency and agility in AI project management.

Managing AI

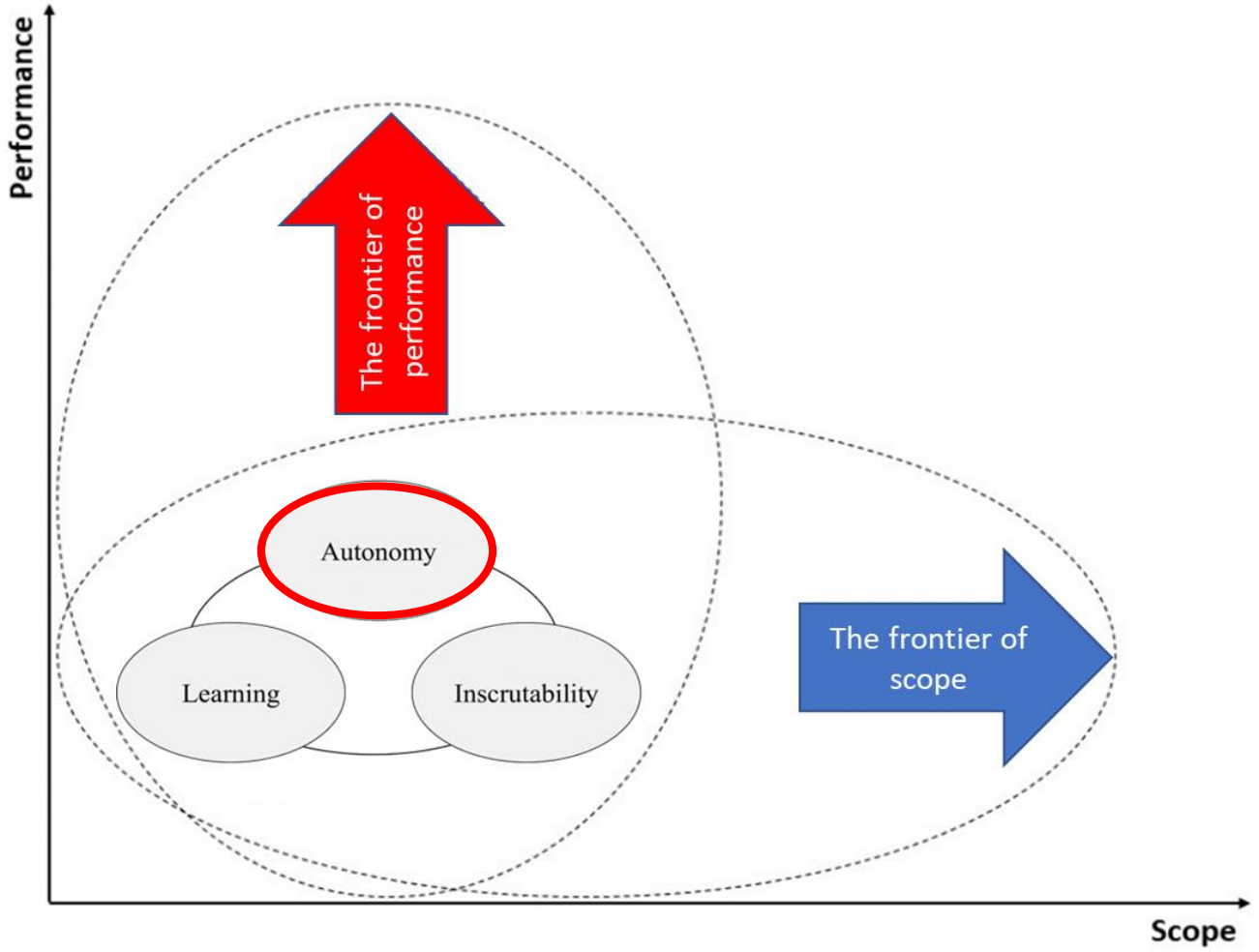
- Managing Artificial Intelligence goes beyond established methodologies like CRISP-DM or other – it extends beyond to push the frontiers of AI itself **[Berente et al. 2021].**

- i.e. push frontier of computational advancements **performance** and **scope**
- This is ensured by conceptualizing three different, interrelated facets of these frontiers:
  - **autonomy**,
  - **learning**,
  - **inscrutability**



Managing Autonomy in AI

- Autonomy refers to the ability of (AI) systems to make decisions and take action without human intervention.
- Examples: Self-driving cars, robo-advisors, AI underwriters.
- Challenges: Ensuring safety, ethical behavior, regulatory compliance.
- Strategies: Establishing control mechanisms, monitoring systems, fail-safe mechanisms.

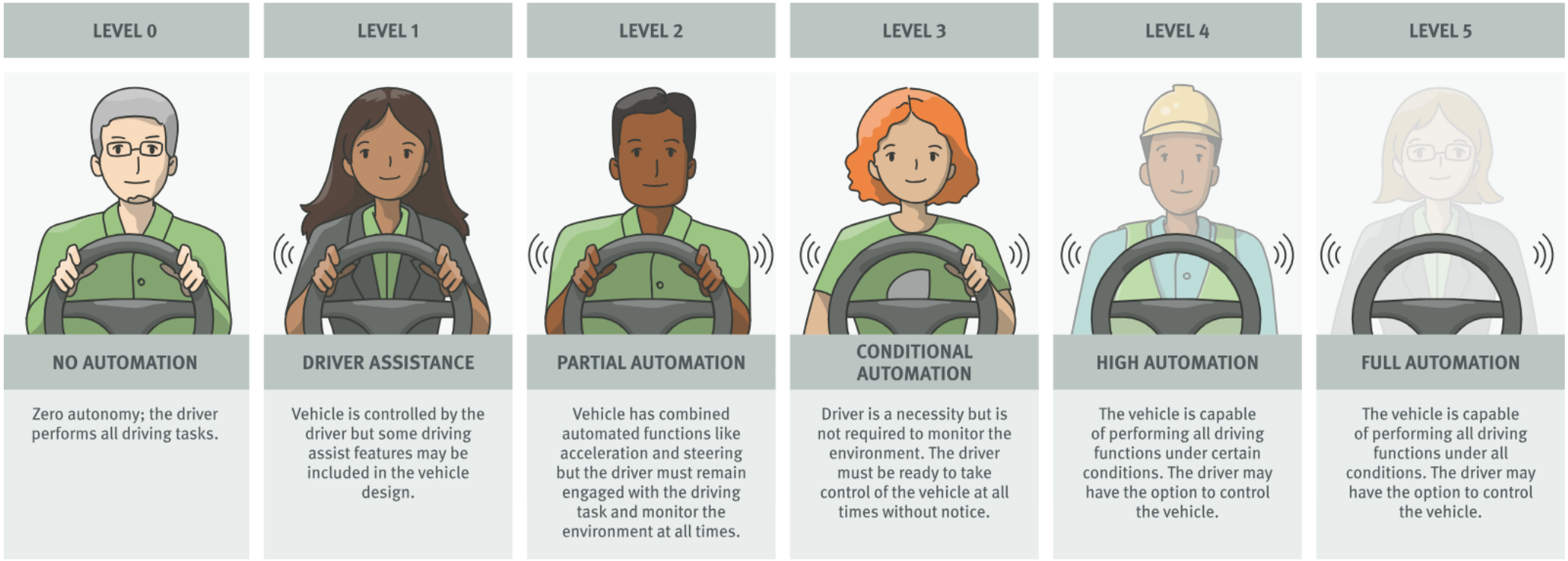


Frontier in Managing Autonomy

1. **Original Frontier**:
   - Human-bracketed AI , where AI systems operate within predefined boundaries and depend on human control and supervision.
2. **Contemporary Frontier**:
   - **Generative agency of AI** explores the potential of AI to go beyond predefined boundaries and actively contribute to creative processes
   - **Conjoined agency between humans and AI**: The interaction between humans and autonomous AI is the key managerial issue of our time. It has historically been put in terms of "**augmentation**".
   - An issue with relying on augmentation is the potential negative dependency effects. As people increasingly rely on AI to enhance their tasks and decision-making, they become more dependent on autonomous tools, especially when facing more complex tasks.
3. **Future Frontier…**

Example on Autonomy

US Society of Automotive Engineer and the National Highway Traffic Safety Administration recognize six levels of autonomous driving capability in cars:



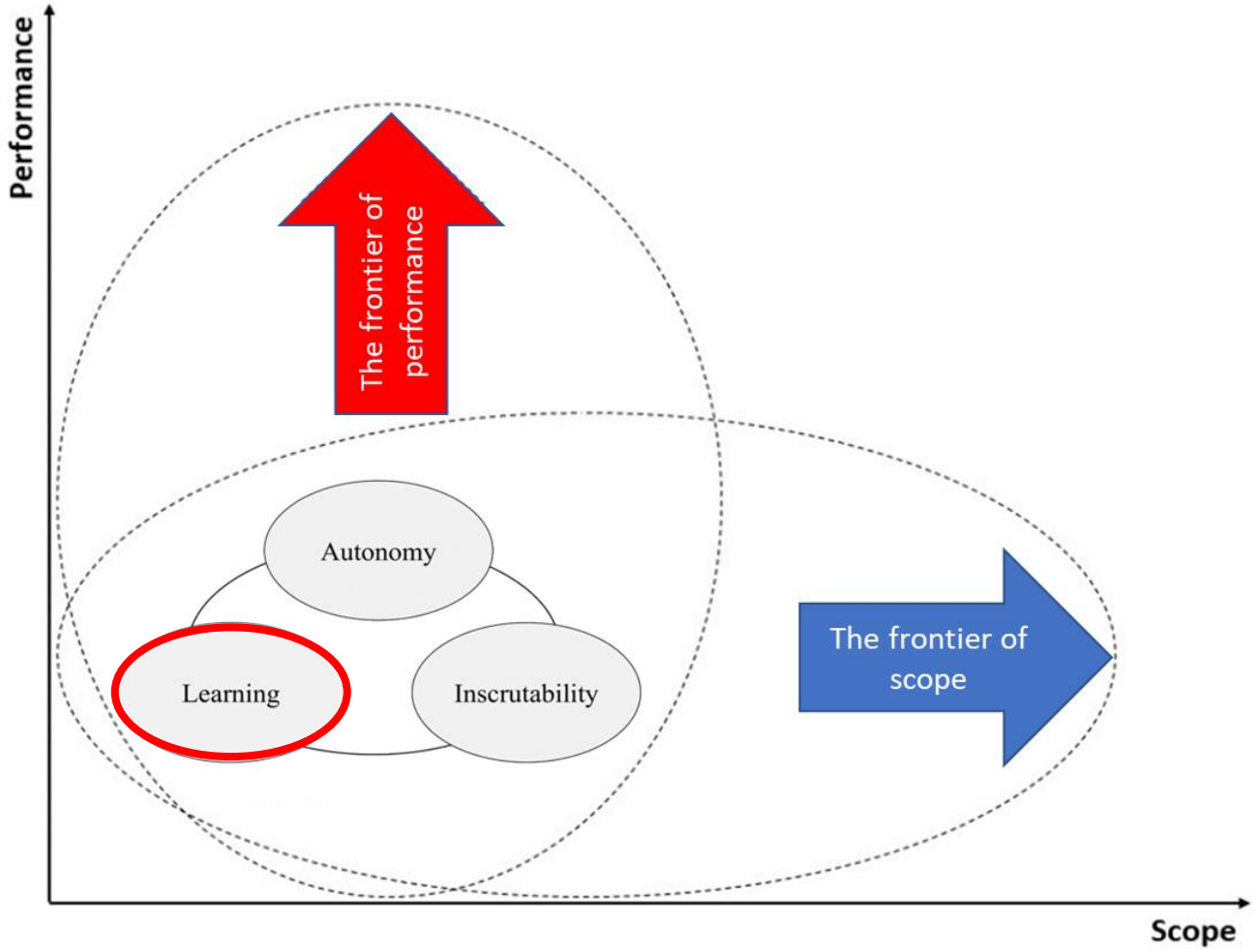
The Society of Automotive Engineers (SAE) levels of automation

- Currently, we are primarily at levels 1 and 2.
- Fully autonomous vehicles at level 5, capable of performing all driving tasks under all conditions without human intervention, are still in the research and development stage and not yet widely deployed.
- While significant advancements have been made in artificial intelligence and autonomous vehicle technology, achieving a high level of complete autonomy in all driving contexts will require further time and development.

Managing Learning in AI

- Learning pertains to the capacity of AI systems to acquire knowledge, improve performance, and adapt to new circumstances through experience.

- Examples: Machine learning, deep learning, reinforcement learning.
- Challenges: Data quality, bias, interpretability of learning outcomes.
- Strategies: Training and validation processes, data governance, algorithm transparency.



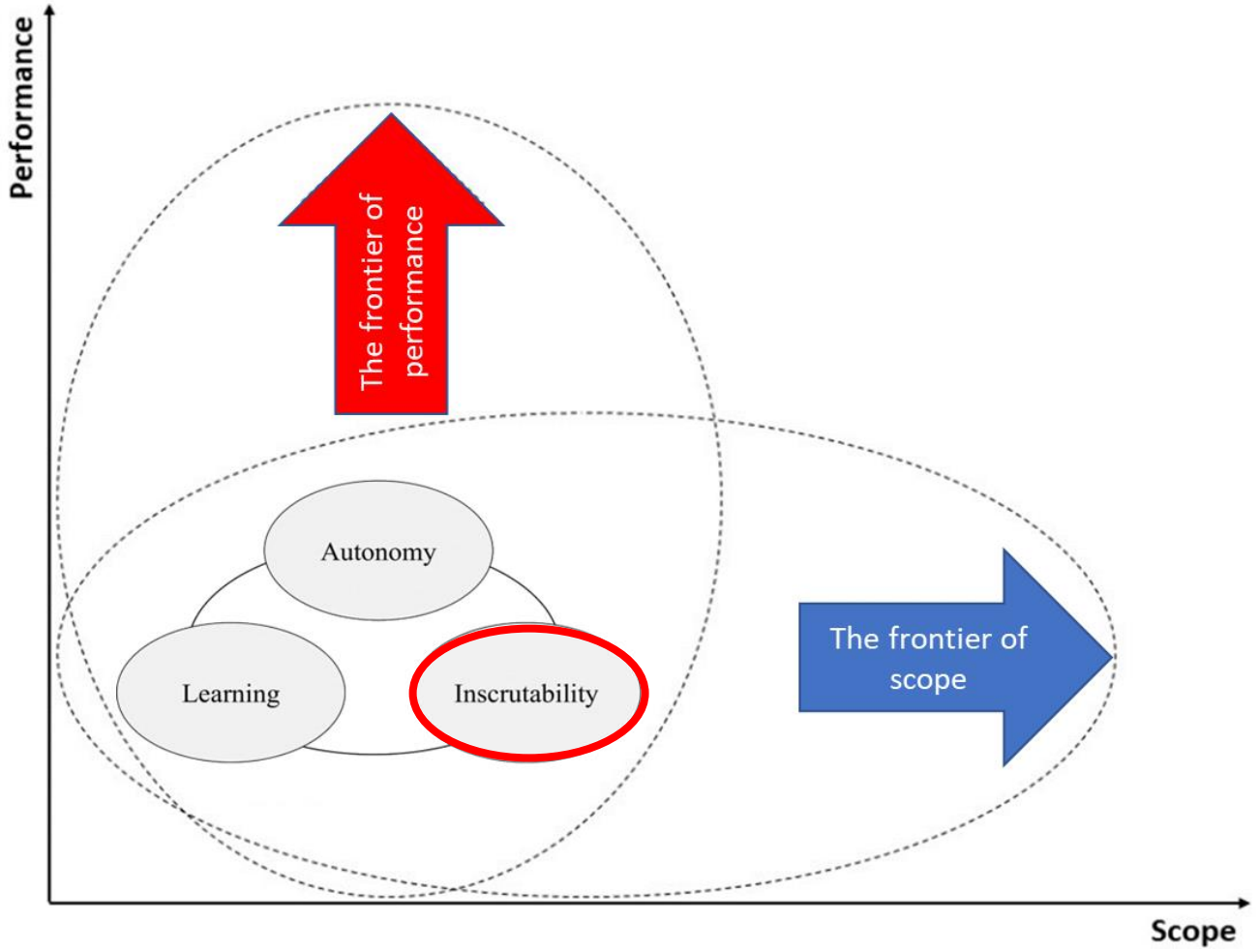
Frontier in Managing Learning

1. **Original Frontiers**:
    - Learning from Structured Proprietary Datasets
    - Human-Driven Data Analysis: crucial role of human involvement in data analysis processes
2. **Contemporary Frontiers**:
    - Large-Scale Trace Data: vast amounts of data generated from various sources, such as social media, online platforms, and IoT devices
    - Human-Unmediated Analysis: This frontier explores the potential of AI to perform analysis tasks without human intervention. It involves developing AI systems that can autonomously analyze data, make decisions, and take actions, minimizing the need for human input or oversight
3. **Future Frontier…**

Managing Inscrutability in AI

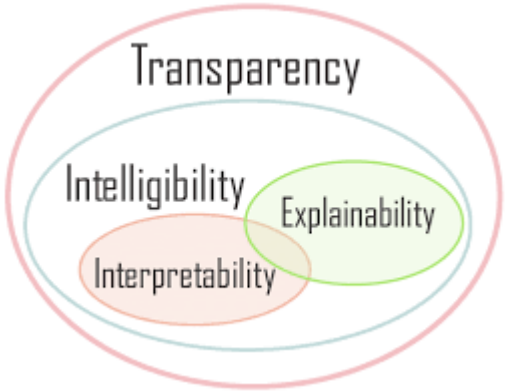
- Inscrutability refers to the challenge of understanding and interpreting the decision-making processes of AI systems.
- As AI becomes more complex and sophisticated, it can become difficult for humans to comprehend the underlying logic and reasoning behind AI-generated outcomes, leading to concerns about transparency, accountability, and potential biases.

- Examples: Black-box models, complex neural networks.
- Challenges: Interpretability, accountability, trust.
- Strategies: Model explainability techniques, interpretability standards, ethical guidelines.



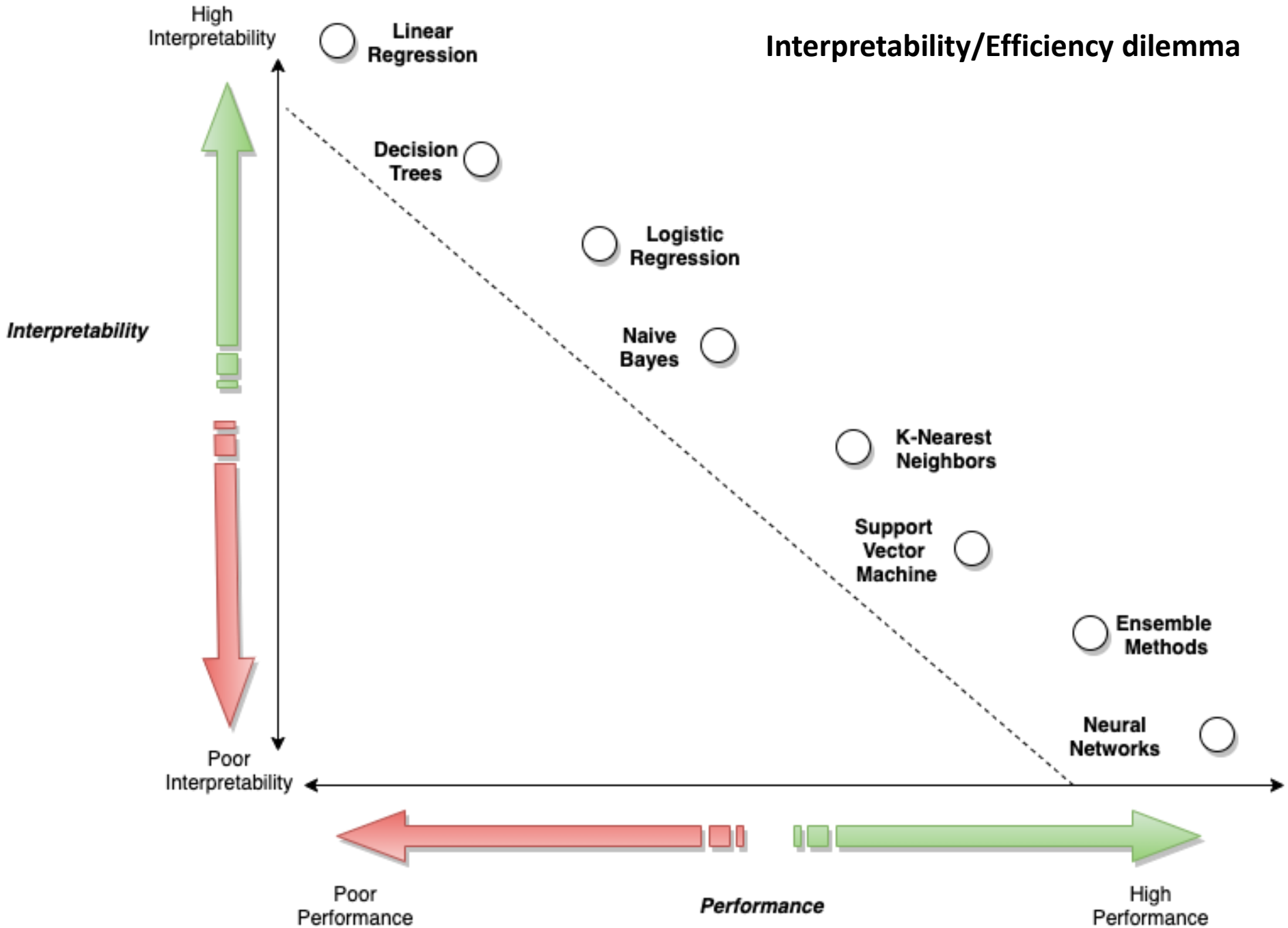
Frontier in Managing Inscrutability

1. **Original Frontiers**:
   - Explicit deterministic logic: This frontier focuses on using explicit rules and logical reasoning that lead to a specific outcome.
   - Manually generated reasoning: relies on human knowledge and expertise to guide the system's actions
2. **Contemporary Frontiers**:
   - Move to non-deterministic (probabilistic, fuzzy etc.) algorithmic logic : In this frontier, AI systems utilize complex algorithms that are not easily understandable or explainable by humans.
   - Self-evolving, genetic deep learning algorithms: This frontier explores the use of deep learning algorithms that can evolve and improve over time through genetic or evolutionary processes.
   - Inscrutability now carries at least four interdependent emphases:
     - ✓ **Opacity**: refers to the lack of visibility into an algorithm (black box)
     - ✓ **Transparency**: involves disclosure and is therefore a strategic management issue.
     - ✓ **Explainability**: refers to an algorithm's ability to be codified and understood at least by some party
     - ✓ **Interpretability**: refers to the understandability and sensemaking on the part of particular humans

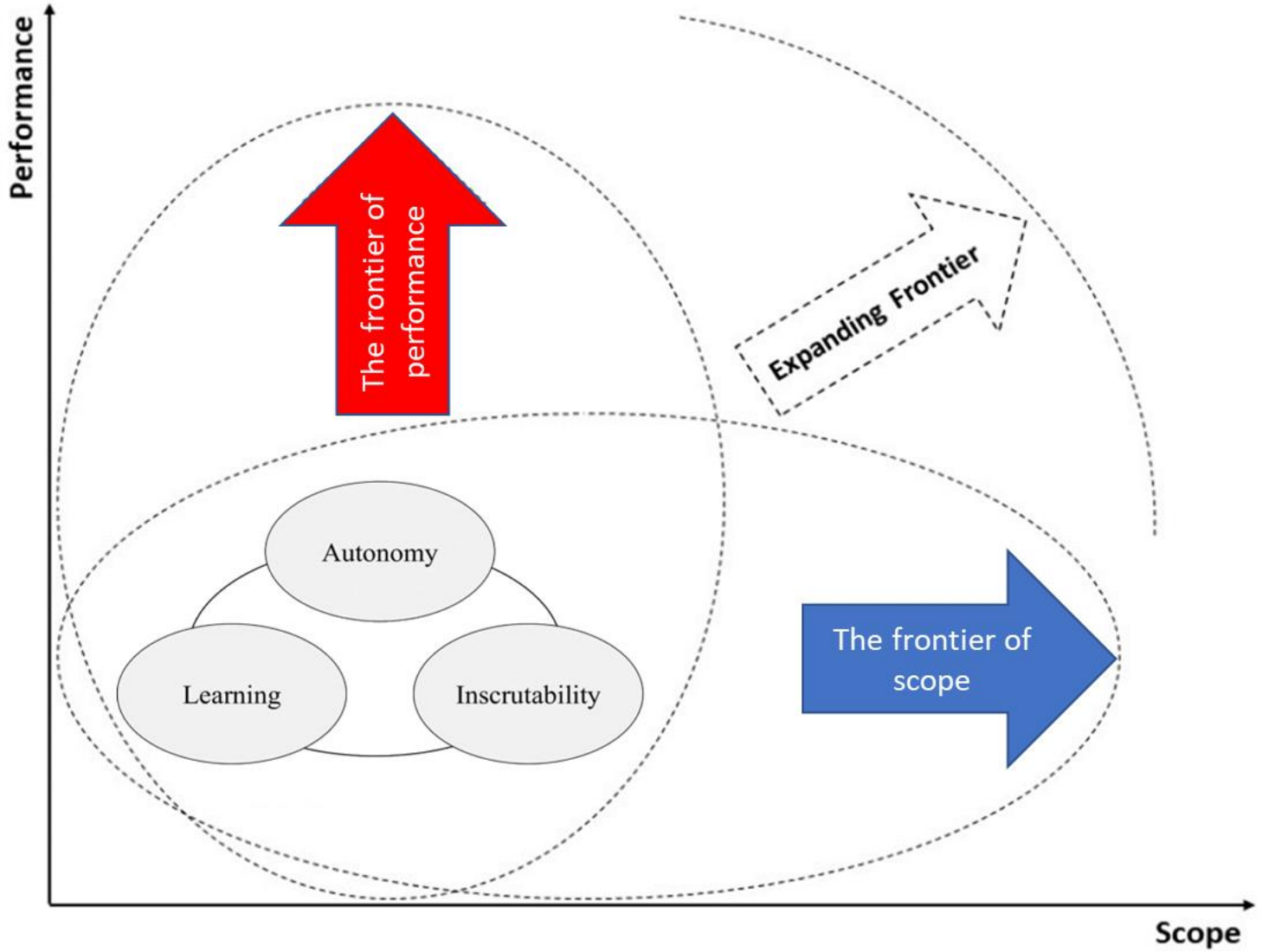


3. Future Frontier…

Interpretability/Efficiency dilemma



Future Frontiers in Managing AI



Future Frontier in Managing AI (1)

1. **Future Frontier in managing Autonomy:**
   - AI and physicality: This emerging frontier highlights AI's ability to autonomously construct and enact physical and tangible aspects of our lived experiences (e.g. Metaverse).
2. **Future Frontier in managing Learning:**
   - Adversarial Learning : In the past few years, AI researchers have developed various techniques to make machine learning models more robust against adversarial attacks. The best-known defense method is 'adversarial training', in which a developer patches vulnerabilities by training the machine learning model on adversarial examples [see AI attacks section].
   - Advances to adversarial learning are accelerating the development and testing of AI applications. One can imagine a future, perhaps powered through quantum computing, where even more development work builds on adversarial learning. This would have several managerial implications.

Future Frontier in Managing AI (2)

3. Future Frontier in managing *Inscrutability*

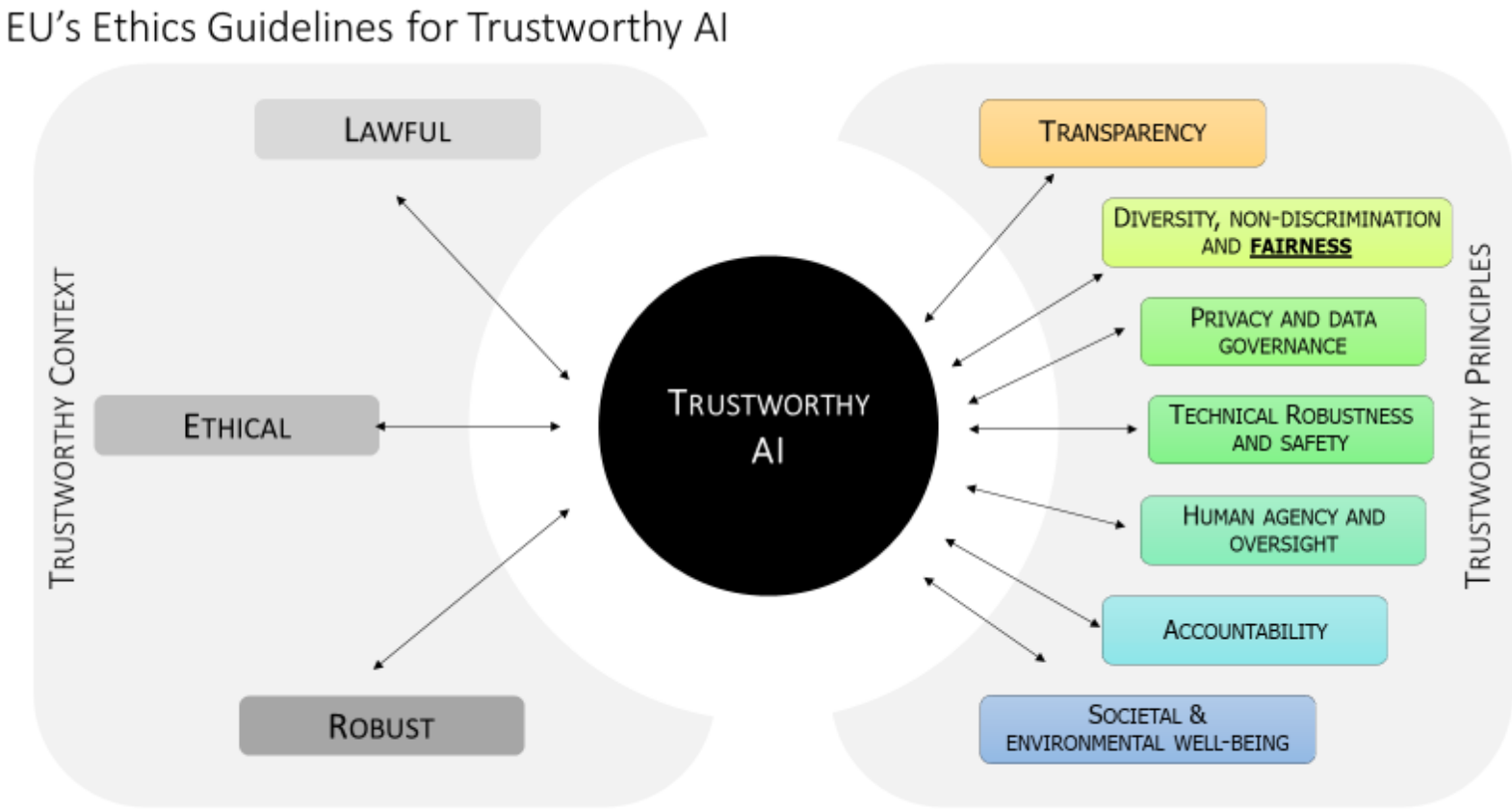
Social context and interpretability: This emerging frontier highlights the importance of considering the social context and interpretability of AI systems. It recognizes that AI decisions cannot be solely based on computational or cognitive factors but must also account for the social dynamics and cultural contexts in which they operate.

Example : Suppose we have an AI system deployed in a healthcare setting. The system is designed to analyze medical records and provide recommendations for personalized treatment plans for patients with a certain medical condition.

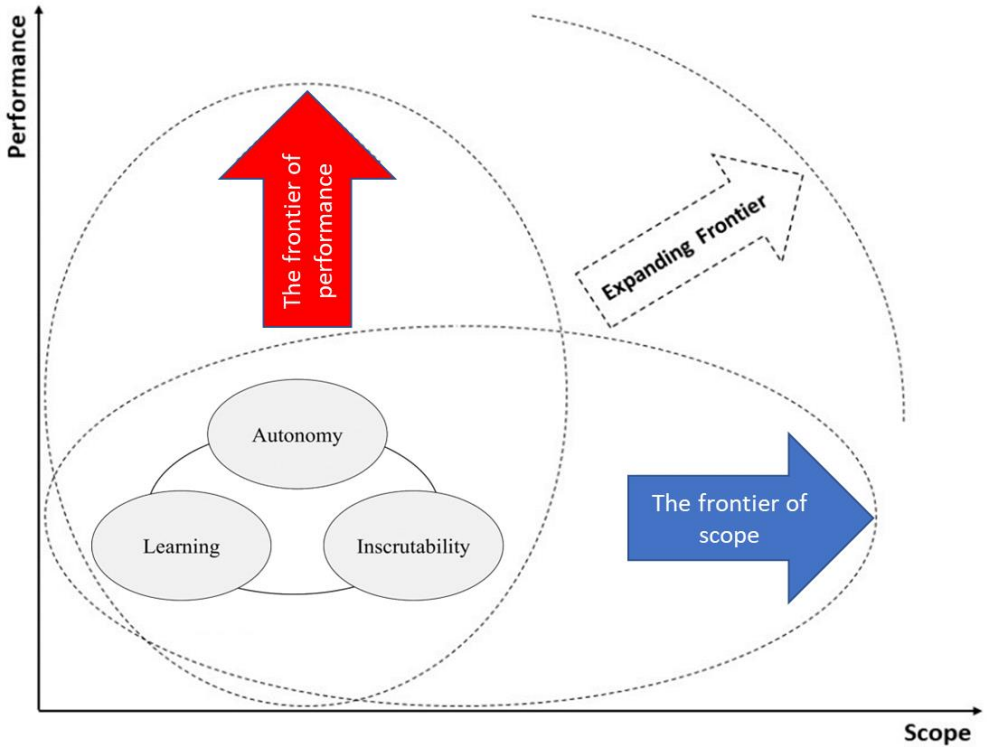
- In this scenario, considering the social context and interpretability of the AI system becomes crucial. The AI system must take into account the unique social and cultural factors that can influence healthcare decisions and outcomes.
- For example, in some cultures, there may be specific beliefs, practices, or preferences regarding healthcare treatments or the involvement of family members in decision-making. The AI system should be sensitive to these cultural nuances and provide recommendations that respect and align with the patient's cultural values and preferences.
- Interpretability also plays a vital role in healthcare AI systems. When the system provides treatment recommendations, it should be able to explain how it arrived at those decisions. This is particularly important in healthcare, where transparency and trust are crucial.

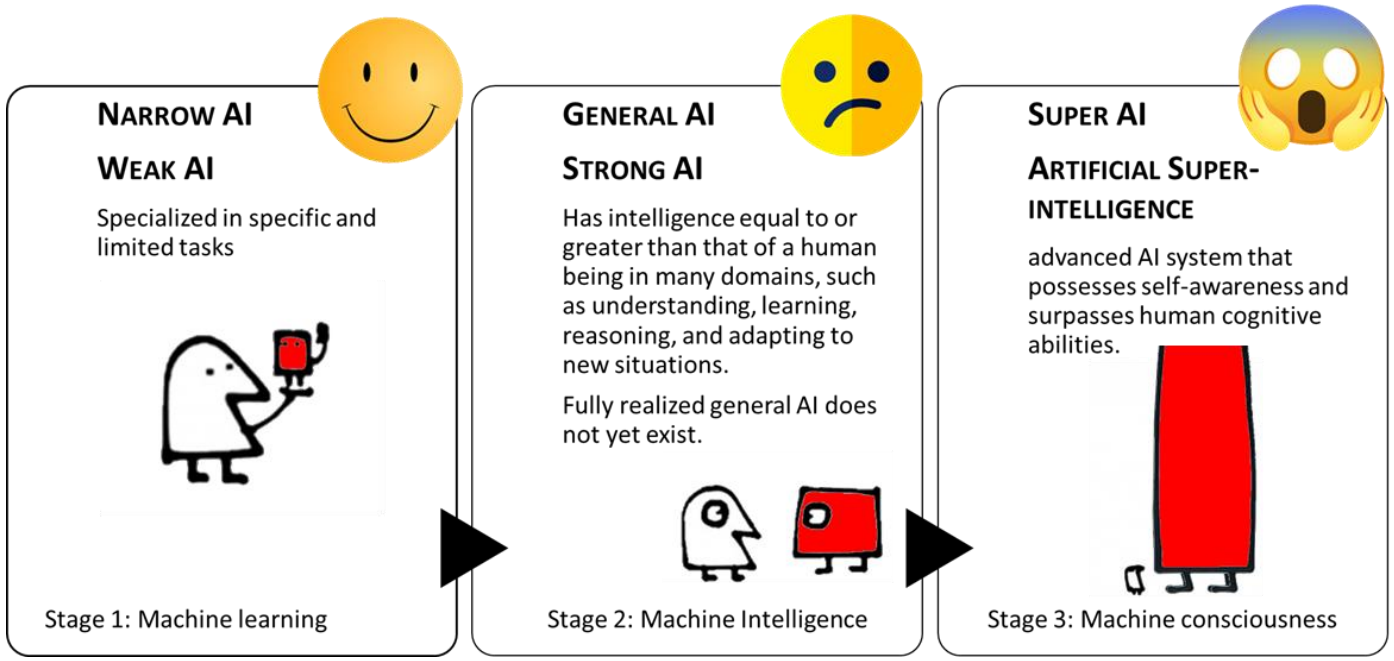
A Future Frontier That Cuts Across All Facets of AI: Ethical Issues

- Recognizing that ethical issues intersect with all facets of AI.
- Addressing concerns regarding automation and its impact on the workforce.
- Navigating data privacy, fairness, justice, discrimination, bias, and legal aspects associated with AI.
- Proactively addressing ethical challenges and responsible approaches to AI development, application, and governance.



Future Frontiers in Managing AI





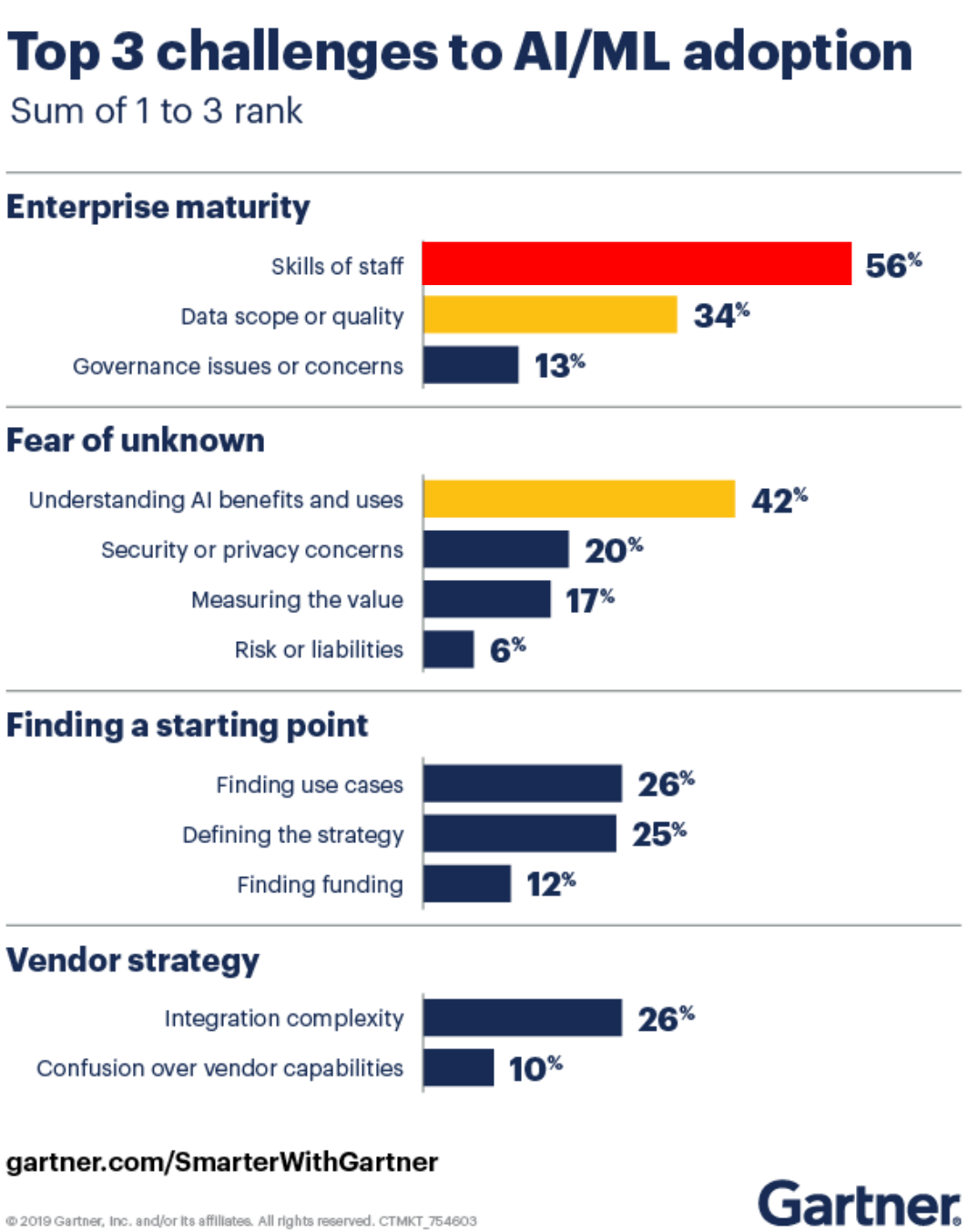
Stages of AI

AI Governance vs Managing AI

- AI governance acts as a **transversal** aspect that spans across all stages of managing AI projects.
- It provides guidelines and principles for the responsible and ethical use of AI technologies.
- AI governance ensures that ethical considerations, legal requirements, and responsible practices are integrated into decision-making processes.
- It helps align project goals with broader considerations, promoting the development and deployment of AI technologies for the benefit of humanity while mitigating potential risks.
- AI governance serves as an essential aspect throughout the different stages of managing AI projects, upholding ethical, legal, and responsible practices, and fostering trust in AI systems.

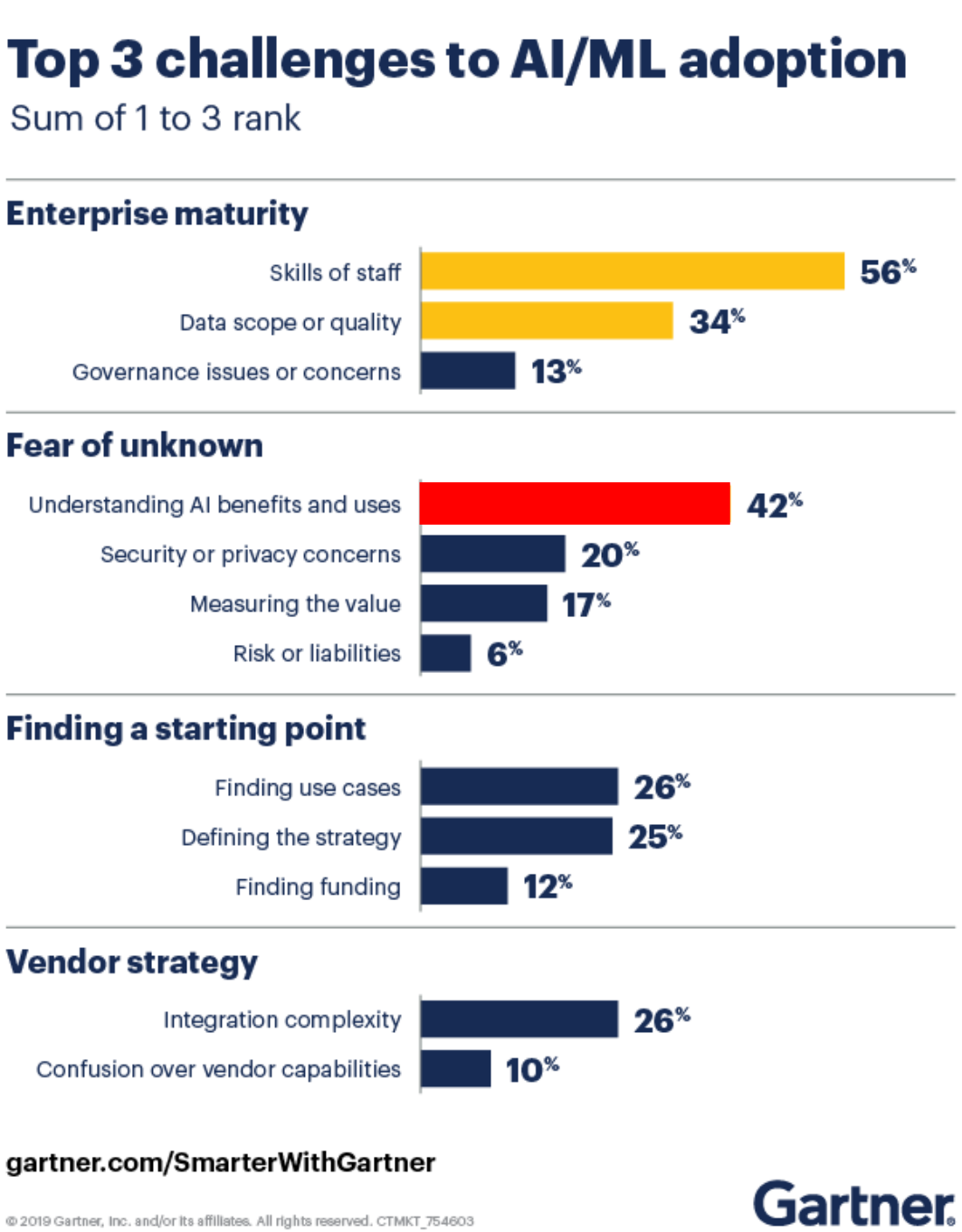
The Challenges of AI Project Development: Why AI Projects Fail ?

1. The first barrier is **skills**. Business and IT leaders acknowledge that AI will change the skills needed to accomplish AI jobs. 56% of respondents said that acquiring new skills will be required to do both existing and newly created jobs, according to a Gartner Research Circle survey.



The Challenges of AI Project Development: Why AI Projects Fail ?

2. The second top challenge is the **fear of the unknown**, which can be attributed to misaligned expectations. 42% of respondents face difficulties in fully understanding the benefits and use of AI in the workplace. Quantifying the advantages of AI projects becomes a major challenge for business and IT leaders. While some benefits, such as increased revenue or time savings, can be well-defined, others, such as improving customer experience, prove difficult to precisely define or accurately measure.



The Challenges of AI Project Development: Why AI Projects Fail ?

3. The third challenge is the **full data scope** **or** **the data quality** derived from AI. Successful AI initiatives depend on a large volume of data from which organizations can draw information about the best response to a situation. Organizations are aware that without sufficient data — or if the situation encountered does not match past data — AI falters. Others know that the more complex the situation, the more likely the situation will not match the AI's existing data, leading to AI failures.