Just in Time

AI Fairness

Addressing Ethical and Reliability Concerns in AI Adoption

Dec 2021

iason
ESSENTIAL SERVICES FOR
FINANCIAL INSTITUTIONS

Executive Summary

During the last 15 years, there has been considerable advancements in AI complexity and performance.

- In many areas, AI achieves comparable or superior results to **human-expert judgment**.

AI emulates a certain behavior by leveraging algorithms designed to accomplish general computational tasks.

- Those algorithms use very different functioning criteria with respect to the logics adopted by human cognition to reach the same goal.

Machine and Deep Learning (ML) are AI core technologies. ML recognizes multivariate mathematical patterns and use them to forecast, classify and more.

- ML internal functioning is **hard to interpret** in human logic.
- Frequently ML consists of **black-box algorithms** which may generate reliability, governance and ethical issues.

Considering the widespread adoption of AI, even in high-risk areas (e.g. creditworthiness), increasing concerns have risen about reliability, resiliency and ethical threads.

At a Glance

01

Social and Ethical Implication

Ethical Implications of AI Adoption

AI Risk Factors

AI Principles and Implications

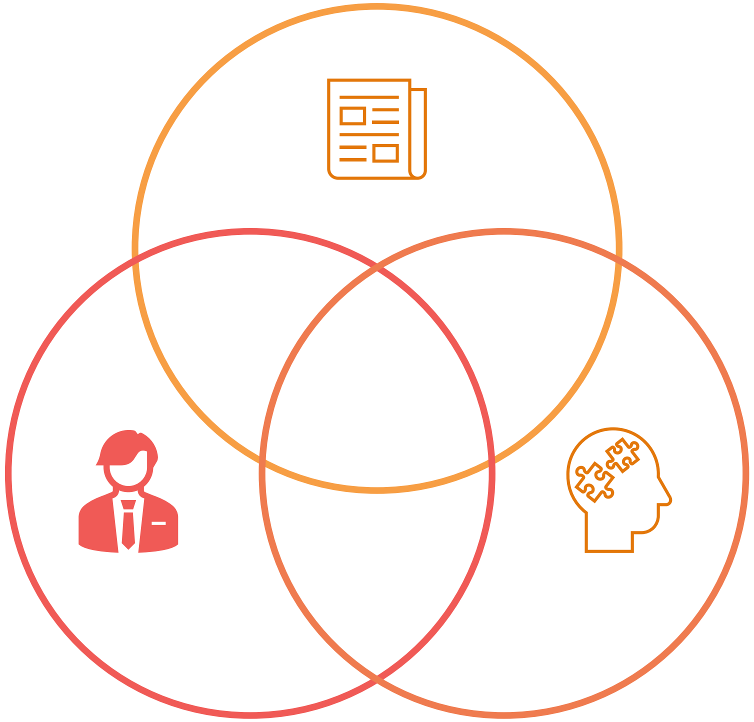
Social and Ethical Implication 1/3

Ethical Implications of AI Adoption

AI technologies have **disruptive impact** on the industrial and economical organization of society but may have inadvertent effects on individuals.

Type of Information

The type of **information we are exposed to** is increasingly determined by AI systems.



Personal Information

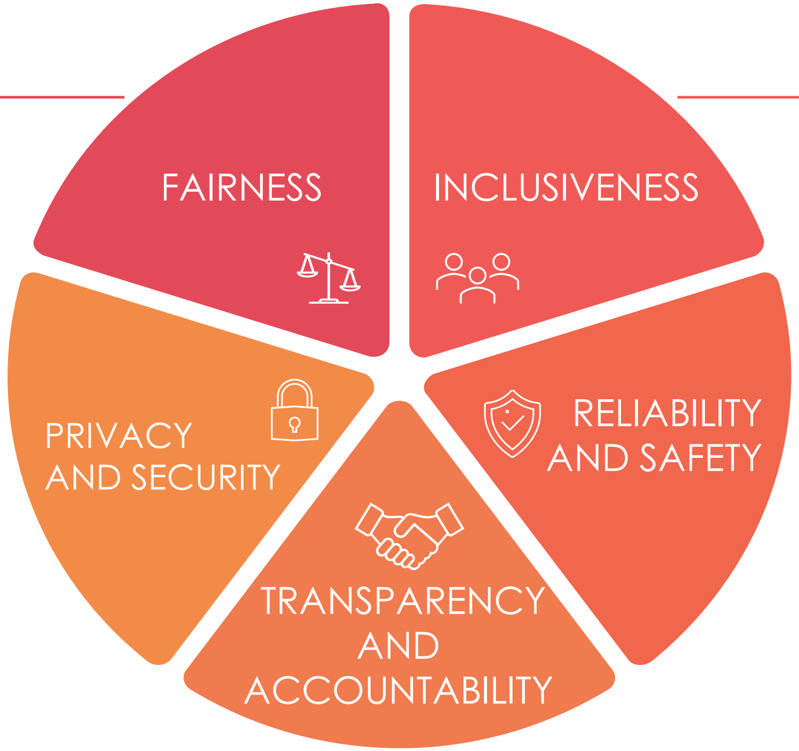
Since AI holds in data consuming algorithms, the use of personal information has been intensified in ways that can intrude on privacy. AI **can infer confidential information**, even when not explicitly exposed.

Ethical Implications

AI can be extremely beneficial for economy and social growth. It may also result in outcomes with serious ethical implications, exposed ranging from **data protection** to **unfair AI adoption** to the expenses of human dignity and rights.

Social and Ethical Implication 2/3

AI Risk Factors



An increasingly important concern as AI systems have the **potential to condition** social, psychological and economic conditions even unintentionally.

Privacy is a fundamental right. AI **consumes data to learn and predict**. Notably, AI also generate data in the form of predictions.

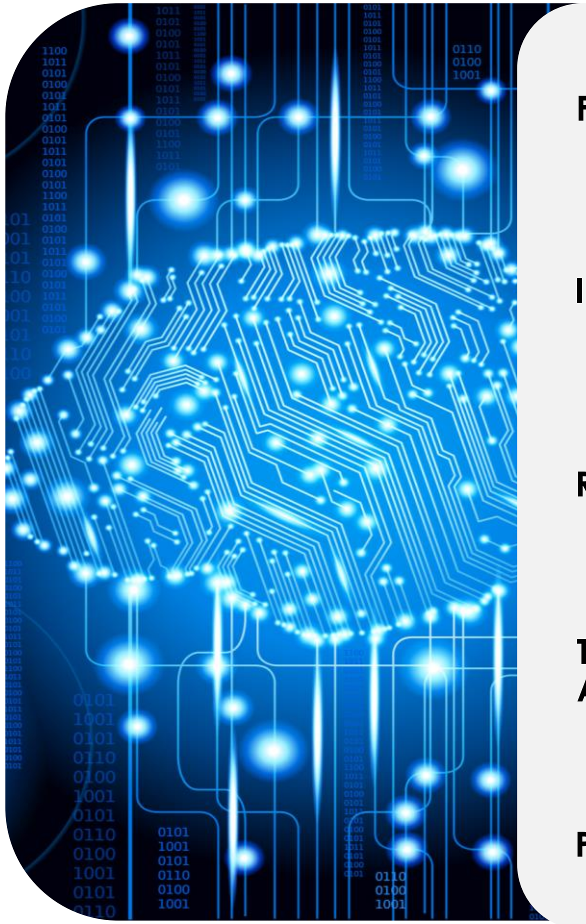
Related to AI systems that are trained to emulate data; data which could mainly **reflect the majority** group.

Related to AI applications **developed and tested on data samples**. For example, systems designed to forecast events leverage past data to predict future outcome.

Accountability refers to the necessity to **justify AI predictions** to users interacting with the system.

Social and Ethical Implication 3/3

AI Principles and Implications



Fairness

1

Biases may arise as **AI reflects a real bias existing in society**, as discrimination or systematic differences across groups. They are present in data used for AI development, as a result of data generation, collection or selection processes. Finally, an AI architecture could be intrinsically biased, optimizing with metrics favoring some groups.

Inclusiveness

2

AI systems may exhibit a **behavior which is not useful for minority groups**. Not equally distributed data access and computational resources for AI applications may cause exclusion of part of society from benefiting from AI

Reliability and Safety

3

There is **no guarantee AI systems can predict an event which has never occurred** in training data. Tiny changes in input data may result in crucial changes in AI behavior. The lack of resembling human logic, makes hard understanding how much an AI system can be trusted.

Transparency and Accountability

4

Transparency includes inspecting model logic and reproducing the dynamics resulting in a certain prediction. It also implies **the need for representation of the moral values** and societal norms holding in the context of operation.

Privacy and Security

5

Various types of data **may rise privacy concerns**. Even if sensible information is anonymized or deleted, AI may be able to infer such information and disclose it or use it to inform predictions.

02

Normative Standards and Guidelines

Regulatory Framework: EC Proposal

Regulatory Framework: Objectives vs Achievements

Regulatory Framework: High Risk, Low Risk, Challenges

Normative Standards and Guidelines 1/3

Regulatory Framework: EC Proposal

The regulatory landscape of AI is an **emerging issue at global level**, and currently it sees involved political entities, as the European Commission, along with supra-national bodies such as IEEE and OECD landscape. The European Commission (EC) published some AI ethical guidelines in 2019 to both encourage AI and manage associated risks. In April 2021, a proposal for a Regulation on Artificial Intelligence was released.

At a high-level EC proposal is organized around three main factors:

Requirement to **conduct assessments of AI risks**;
- Including to document how risks have been addressed;

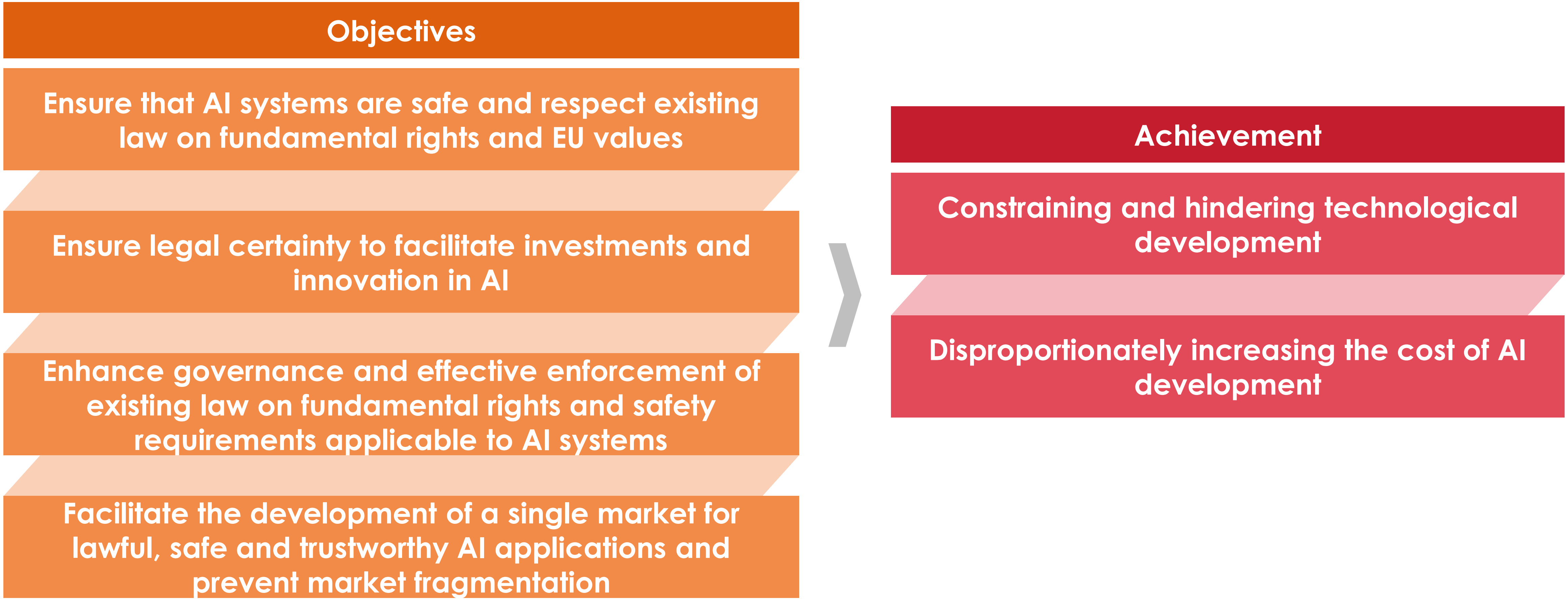
Accountability and independence of regulatory entities;
- Determine liabilities and conflicts of interest;

Continuous **monitoring** of AI algorithms;

Normative Standards and Guidelines 2/3

Regulatory Framework: Objectives vs Achievements

The following specific objectives are pursued by the regulation with regard to the use of AI in EU.



Objectives

- Ensure that AI systems are safe and respect existing law on fundamental rights and EU values
- Ensure legal certainty to facilitate investments and innovation in AI
- Enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems
- Facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation

Achievement

- Constraining and hindering technological development
- Disproportionately increasing the cost of AI development

Normative Standards and Guidelines 3/3

Regulatory Framework: High Risk, Low Risk, Challenges

European Commission classifies AI applications on their **inherent risk level** and set clear requirements for AI systems for high-risk and limited-risk systems. The AI Regulation also imposes **obligations on users of high-risk AI systems** for instructions and monitoring purpose.

High-risk Systems concerns the ones related to **safety components** regarding environments characterized by factors threating security.

Limited-risk Systems concern about **users' awareness** of dealing with AI system.

03

Limits of Data-driven Approaches

AI Underneath the Hood

Machine Learning Burden and Delight

Prediction Degradation After Distributional Shift

Limits of Data-driven Approaches 1/3

AI Underneath the Hood

AI Underneath the Hood

Machine Learning (ML) algorithms are problem-agnostic and learn to emulate a phenomenon without necessarily grasping the underlying dynamics.
We claim pure data-driven use of those algorithms be intrinsically a poor approach, leading to enormous technical and operational risks.

1 ML encompasses several data-driven models which learn from data to emulate a behavior without being explicitly programmed to do so

2 ML is the core technology in AI, but other tools are inherited from linguistics, computer vision, search engines, inferential statistics and more

3 AI performance depends on data availability and computational power

4 The explosion of data generated by modern devices, along with the fall of computing power costs, has contributed to sticking AI improvements

5 Many believes that optimal decisions can be made by complex algorithms in a purely data-driven way. But this is far from true.

Limits of Data-driven Approaches 2/3

Machine Learning Burden and Delight

Optimal inference

Optimal inferences on *y* given observations *X* requires computing the full joint distribution $P(y, x_0, x_1, \ldots, x_n)$. But in most cases, such a distribution cannot be computed

Functioning logics

ML predictions do not respond to human or formal logic, nor “reason” on the underlying dynamics of the phenomenon being modelled

Mathematical tricks

Part of ML power comes from using mathematical tricks to approximate $P(y \mid x_0, x_1, x_n)$ without computing the full joint distribution $P(y, x_0, x_1, x_n)$

Generalization limits

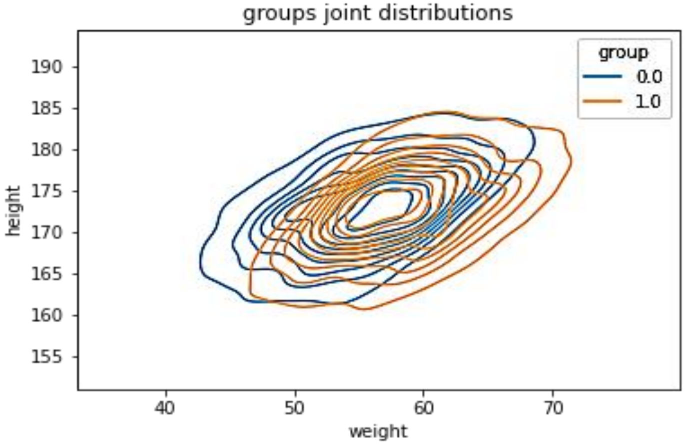
Some ML algorithms (e.g. decision trees) suffer considerable generalization limits and poor resiliency to small input data changes

Limits of Data-driven Approaches 3/3

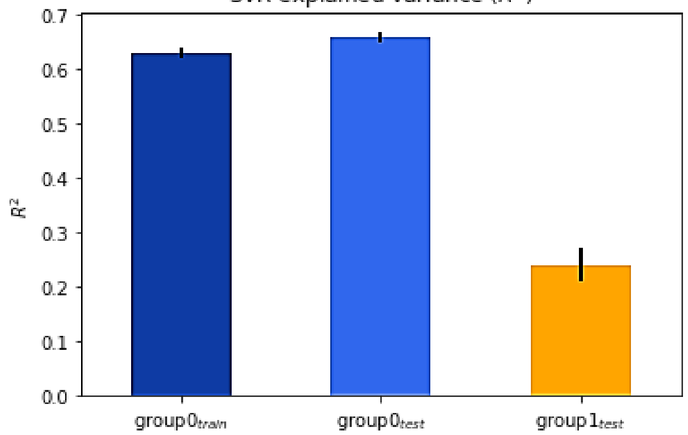
Prediction Degradation After Distributional Shift

A simple experiment (predicting height from weight) reveals a small shift between train/test populations results in considerable inferential power loss.

groups joint distributions



SVR explained variance (R^2)



SVR Regression

A Support Vector Machine Regressor (SVR) is trained on group0 data to predict height from weight and gender. The resulting model can explain more than 60% of height variability from weight on an independent dataset (group0 test).

Rule out Overfitting

Notice, predictions on test (grou0test) does not significantly differ from train (group0train) excluding any overfitting effects, which may impair generalization.

Prediction Failure

However, when the model is assessed on a population with slightly greater man proportion (group1test), the model losses dramatically drops its predictive power.

04

Alternative Modeling Approaches

Probabilistic Programs

Modeling as Simulation

Just in Time

Alternative Modeling Approaches 1/2

Probabilistic Programs

AI applications should be designed in a thoughtful manner, by considering all the available knowledge on the process being modelled.

Model-based reasoning

Theoretical knowledge can be encoded in a graph illustrating the relationships expected to appear in data

Probabilistic computational graph infer statistical conclusions from complex combinations of aleatory observations

Statistical conclusions

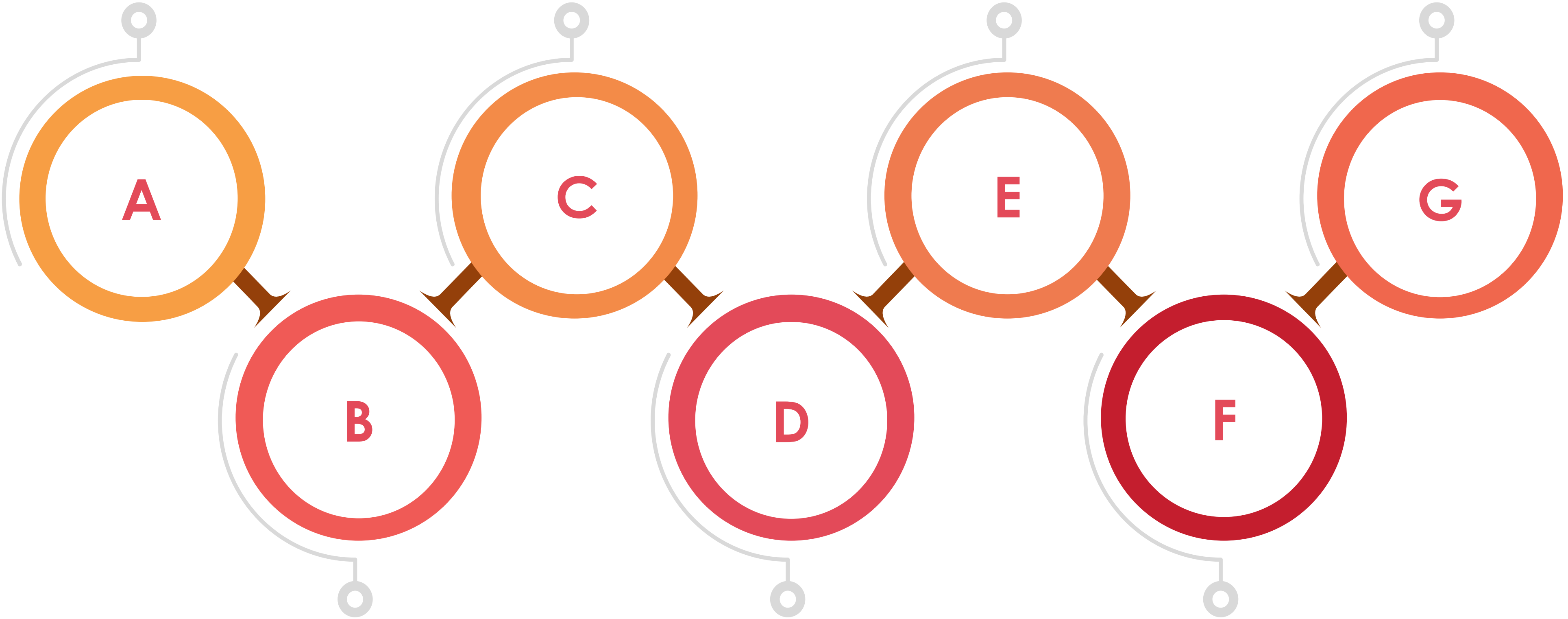
Theoretical knowledge can be encoded in a graph illustrating the expected relationships

Counterfactual reasoning

In contrast with ML, probabilistic models allow to run inference under conditions not represented in training data

Alternative Modeling Approaches 2/2

Modeling as Simulation



A: Bayesian inference proceeds by building a generative model of the data (simulator)

B: Model parameter consistent with observations are estimated via MLE and Bayesian rule

C: Posterior distributions for target and parameters are obtained

D: Bayesian networks allow backward inference (**nonmonotonic reasoning**), by updating the parameter distributions when information enter any nodes

E: Bayesian networks can handle incomplete information, while ML requires complete features

F: The model can generate data consistent with any nodes set to a specific value (**counterfactual reasoning**)

G: Bayesian networks can explore all the possible outcomes produced by different circumstances

05

Fairness Definition

Fairness Definition 1/2

A model is considered fair, in relation to protected groups, if errors are distributed similarly across protected groups. However, there are many possible ways to interpret this. Below the most used fairness metrics are listed.

Demographic Parity

Suggests that a predictor is unbiased if the prediction yHat is independent of the protected attribute A

$$\Pr(\hat{y}|A) = \Pr(\hat{y})$$

Notably, such a metric may conflict with equality of opportunity, which allows classification results in aggregate to depend on sensitive attributes, but does not permit classification results for certain specified ground-truth labels to depend on sensitive attributes.

Equality of Odds

A predictor yHat satisfies equalized odds with respect to protected attribute A and outcome Y, if yHat and A are independent conditional on Y

$$\Pr\left(\hat{Y} = 1 \middle| A = 0, Y = y\right) = \Pr\left(\hat{Y} = 1 \middle| A = 0, Y = y\right), y \in \{0,1\}$$

Equalized odds enforces that the accuracy is equally high for all groups and punishes models that perform well on the majority group only.

Fairness Definition 2/2

Just in Time

Some definitions of fairness are mutually incompatible and cannot be satisfied simultaneously

Predictive Parity

A classifier satisfies Predictive Parity if both protected and unprotected groups have equal Positive Predictive Value (PPV: the fraction of positive cases correctly predicted to be in the positive class out of all predicted positive case).

Equality of Odds

A binary predictor $\hat{y}$ satisfies equal opportunity with respect to A and Y if:

$$\Pr\left(\hat{Y}=1\middle|A=0,Y=1\right)=\Pr\left(\hat{Y}=1\middle|A=1,Y=1\right)$$

Equal opportunity is a weaker, though still interesting, notion of non-discrimination, which provides more flexibility.

The described metrics can reveal biases, but do not help in removing them once the algorithm has already been trained. Approaches to handle biases have been developed for all stages of development: data collection (Identify lack of examples or covariates), pre-processing, in-processing and post-processing.

06

Bias Mitigation

Bias Mitigation 1/4

During Pre-processing

Straightforward approach for eliminating biases from datasets consists in removing the protected attribute and other elements of the data that are suspected to contain biasing information. Unfortunately, such suppression rarely suffices. There are often subtle correlations in data leading the algorithm to infer the protected attribute. The degree to which there are dependencies between data X and the protected attribute p can be measured using Mutual Information. Such dependency is known as latent prejudice. As this measure increases, the protected attribute becomes more predictable from the data.

1 **Re-weighting data pairs**: re-weighted the {x, y} pairs in the training dataset so that the existing cases where the protected attribute p is linked to a positive outcome in the disadvantaged group are more highly weighted. Then a classifier is trained considering these weights in the loss function. Alternately, re-sampling the training data according to these weights and using a standard classifier.

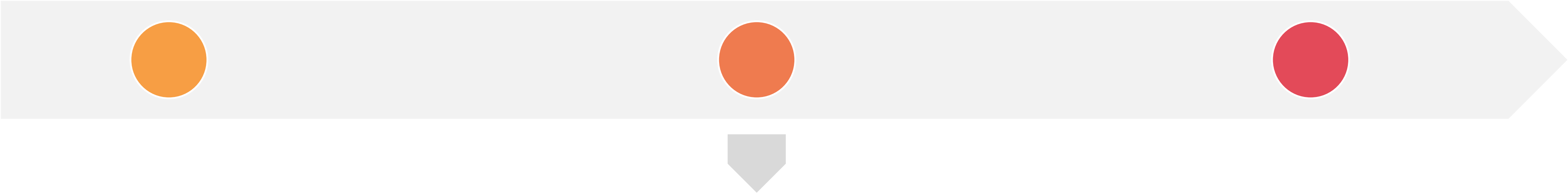
2 **Manipulating observed data**: individual features x from data X can be manipulated in a way that depends on the protected attribute p. Each dimension x is divided in case where the protected attribute p is 0 or 1. Then the cumulative distributions F0[x] and F1[x] are computed and aligned to a median cumulative distribution Fm[x].

3 **Manipulating labels and data**: find a randomized transformation Pr(x',y'|x,y,p) transforms data pairs {x,y} to new data values {x',y'} in a way that depends explicitly on the protected attribute p. Such a problem is formulated as an optimization problem in which prejudice is minimized with constraints on the level of distortion of the original values. This approach has the advantage of considering interactions between data dimensions.

4 **Manipulating labels**: trained a classifier, then found examples close to the decision surface. They then swapped the labels for some of the edge cases in such a way that a positive outcome for the disadvantaged group is more likely and re-train. This heuristic approach empirically improves fairness at the cost of accuracy.

Bias Mitigation 2/4

Pre-processing



During In-processing

An elegant approach for removing bias during training is to explicitly remove such a dependency by leveraging Adversarial Learning. Other easier approaches include penalizing the Mutual Information between p and the prediction using regularization techniques and fitting the model under the constraint that it is not biased.

1 **Prejudice removal by regularization**: proposed adding an extra regularization condition to a classifier for minimizing the Mutual Information between the protected attribute p.

2 **Hyper-parametric model optimization**: Bayesian optimization is chosen to identify which hyper-parameters of the models can identify protected elements of the dataset. Bayesian optimization also allows to optimize multiple metrics simultaneously, leading the model to behave in the fairest way; however, it comes at the cost of very high computational demand.

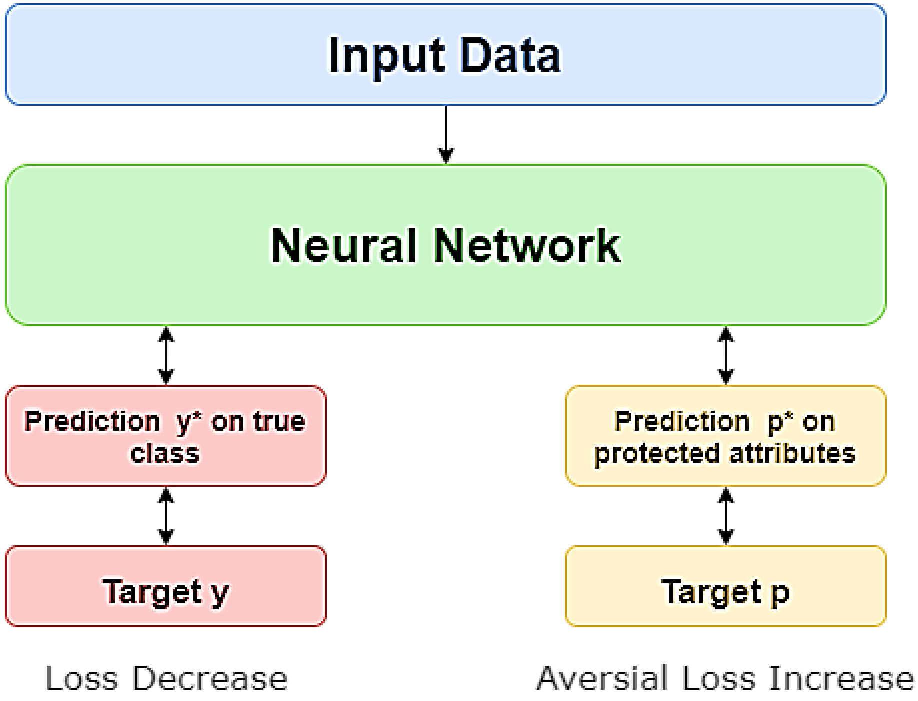
3 **Transfer learning technique**: consist in training the algorithm on data not affected to critical elements and then continuing the training introducing biased information. This technique can only be performed with stochastic gradient algorithms.

Bias Mitigation 3/4

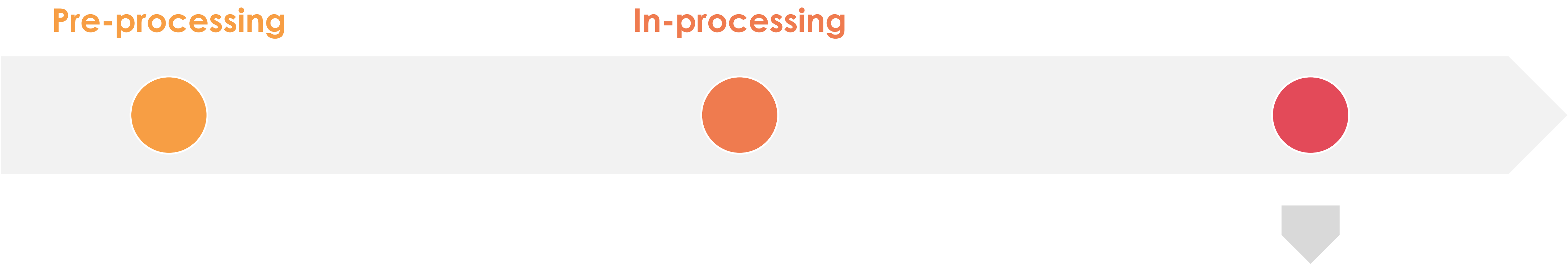
Just in Time

Adversial Learning

Evidence of protected attributes in predictions are reduced by constraining the algorithm to predict and simultaneously fool a second classifier, which in turn attempts to classify the protected attribute p from the output of the first algorithm. Figure illustrates the Adversarial de-biasing mechanism.



Bias Mitigation 4/4



During Post-Processing

Post-processing techniques for addressing fairness provide several practical advantages. The most important benefit is that those techniques do not intervene on any stages of the training process, therefore are de facto the only intervention which can be applied to industrialized application, without affecting the whole pipeline. Similarly, post-processing algorithms do not need access to model functioning, as limit to compute a post hoc mitigated decisions. This is achieved by transforming the model prediction yHat to enforce a specified fairness constraint. Therefore, post-processing approaches provide great flexibility and do not require model retraining.

Company Profile

Iason is an international firm that consults Financial Institutions on Risk Management. Iason integrates deep industry knowledge with specialised expertise in Market, Liquidity, Funding, Credit and Counterparty Risk, in Organisational Set-Up and in Strategic Planning.

Danilo Rubicondo

✉ in

Luca Rosato

✉ in

www.iasonltd.com