Research Paper Series

AI Fairness Addressing Ethical and Reliability Concerns in AI Adoption

Danilo Rubicondo

Luca Rosato

March 2022

iason

ESSENTIAL SERVICES FOR FINANCIAL INSTITUTIONS

TECHNESTHAI

iason
Research Paper Series
Year 2022 - Issue Number 45

Last published issues are available online:
http://www.iasonltd.com/research

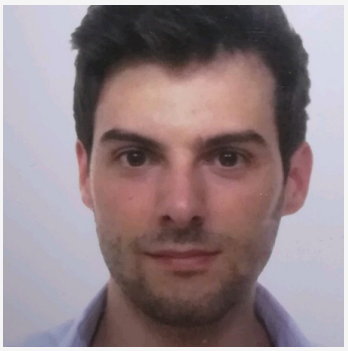
Front Cover: **Piero Dorazio**, *Angolo Blu*, 2000.

Executive Summary

During the last 15 years there has been considerable advancements in AI algorithms, leading AI to achieve comparable - and sometimes superior - performance to human expert judgment. Not surprisingly, AI is now ubiquitous in financial industry: financial firms apply AI to select investments, banks to score creditworthiness, insurances to identify frauds - just to mention a few examples. However, AI algorithms do not resemble (human) logic, but leverage subtle statistical associations in data to make point estimate predictions. This results in black-box architectures, which lack of transparency and do not ensure that predictions generalise to changes in data distribution induced by exogenous factors. Therefore, increasing concerns have risen on AI reliability, resiliency along with several ethical threads, which extend well beyond AI failure or malicious use. In the present document, AI-related risks and their technical origins are discussed. Furthermore, a summary of the regulatory framework for AI, recently proposed by the European Commission, is provided, accompanied by methodologies to address the mentioned issues. Finally, while most AI applications are built on top of data-driven pattern recognition algorithms, which enforce agnostic models on data, the authors advocate the reversed approach. Probabilistic graphs, which can integrate theory-driven models with high-level information, allow to generate predictions in the form of posterior probability distributions. Probabilistic graphs bring multiple technical and epistemological benefits, as exposing their functioning principle, allowing risk assessment, counterfactual reasoning and increasing knowledge of the phenomenon being studied.

About the Authors

Danilo Rubicondo:
Senior Data Scientist
Neuroscientist with over 10 years of experience in the computational modelling of cognitive functions. He is currently working on AI systems supporting credit risk decisions and investment strategies.



Luca Rosato:
Data Scientist
Computer Scientist with experience in AI-Fairness related research. He is currently working on AI systems supporting credit risk decisions and fairness assessment.

Table of Content

AI Fairness Addressing Ethical and Reliability Concerns in AI Adoption

Danilo Rubicondo Luca Rosato

During the last 15 years there has been considerable advancements in AI algorithms complexity and performance. Machine Learning (ML) is the cornerstone of AI, as it allows programs to operate tasks for which they are not explicitly programmed. In other words, ML algorithms can learn to emulate a behavior as exhibited in data. For example, these technologies, in principle, may recognize patterns in the historical daily prices of a stock to predict its future price, without requiring any knowledge of financial markets, technical analysis, and so on. As most ML algorithms are intrinsically mathematical or statistical models, an important limitation is they require a vector of numbers as input data. While transforming input data into a vector is relatively straightforward for many applications, such as stock price prediction, it is very limiting in areas such as Natural Language Processing (NLP) or Computer Vision, where inputs are complex unstructured data, as a newspaper page or a video record. The rise of Deep Learning (DL) enormously boosted AI. Deep Learning algorithms can be viewed as arbitrary sequences (or networks) of ML algorithms, where the model at a certain stage processes the output(s) of the previous one(s). An important advantage of DL is the possibility to take as input arrays of any shape (not just vectors), or even sequences of arrays. This allows to feed images, video, text, sequences of tables with minimal data processing. For example, DL can process time-series information, as historical stock prices and volumes, along with textual information, as news and balance sheet content - and potentially even communication networks - to extract a signal useful for future returns prediction. However, power and flexibility come at the cost of loosing control on how a DL algorithm connects inputs (e.g. news text and past prices) to outputs (e.g. future predictions).

This problem is compounded by the fact ML/DL algorithms (hereinafter we refer to both as ML) do not resemble human reasoning (or formal logic). The connections between inputs X and outcomes y is learnt on the basis of subtle associations (patterns) appearing in training data (X, y). ML learns those specific associations, without necessarily individuating the real dynamic causing y from X. Although some techniques are applied to assert an algorithm can generalise beyond the specific data observed during the learning phase (i.e. training), the risk the algorithm exploits a subtle statistical bias to make predictions is very high. This problem highlights several reliability and ethical dangers, because it may cause not just functioning failures (e.g. prediction errors), but also discriminatory behaviors (e.g. errors in prediction specific to a group). Those issues are discussed in greater detail in the following sections.

Section 2 describes most of the social and ethical implications of the use of AI technologies. Then, Section 3 resumes the current normative standards and guidelines concerned AI governance. Section 4 discusses the technical origin of the issues previously mentioned, with particular emphasis on reliability concerns. Section 5 advocates alternative methodological and modelling procedures to address those limits while also providing considerable epistemological advantages. Finally Sections 6 walks through a technical definition of *fairness* and introduces practical methods to detect and tackle unfairness in AI applications.

1. Social and Ethical Implications

According to a recent update of the International Data Corporation (IDC) Worldwide Semiannual Cognitive Artificial Intelligence Systems Spending Guide, spending on AI systems will reach $77.6 billion in 2022. Notably, this amount is more than three times the $24.0 billion forecast for 2018. AI has undoubtedly benefited to society because it automatises mechanical and possibly harmful tasks, improves production, increases services availability, while creating at the same time new business opportunities and reducing operational costs. Although AI has frequently been accused of leading to a job apocalypse, anecdotal evidence and several studies suggest that human jobs change but donât disappear. For example, a study examined impact of AI and robotics on three prominent sectors: manufacturing, retail banking, and nursing homes. The study found the introduction of AI initially replaced human workers, but often those industries generated new jobs that at least partially offset the losses (Lee, 2020). Indeed, implications of such a widespread AI adoption are more subtle than apocalyptic fears grabbing newspaper headlines.

While malicious use of AI can already be appreciated on a large-scale (see Facebook's news-feed algorithm influencing elections), we believe that criminal abuses are well balanced by uses of AI in countering frauds, crimes and several types of threads. The present document focuses on unintended and unforeseen consequences of AI. In our view, unintended risks may arise because of three factors: i) prediction biases favourable or unfavourable to specific social categories, ii) sudden and unpredictable AI failures in response to environmental changes and iii) unequal availability of AI benefits in economy and society. In the following subsections five main risks deriving from those factors are discussed: fairness, inclusiveness, reliability and safety, privacy and security, transparency and accountability.

1.1 Fairness

AI decisions are informative in critical domains such as credit scoring, employment and medicine. However, those decisions may be biased. In relation to fairness, a bias is a systematic evaluation tendency, which may not necessarily affect prediction performance, but does make unethical discrimination among social groups. Therefore, biases generate unfair decisions. Formally speaking, biases can be conceived as statistical differences among groups, and may arise at three levels. First, biases exist in the real world and AI is guaranteed to detected and exploit them for making inferences. For example, if a certain ethnic group is statistically more exposed to default risk, any ML algorithm would naturally reduce (but not zero) the likelihood of granting a loan to any individuals of that group. Second, even if biases do not exist in the real world, they may appear in data. This could happen for a variety of reasons ranging from reflecting direct or indirect discriminatory beliefs to mere chance of sampling error. Third, biases may arise from technical choices in the AI architectures, such as loss functions penalizing extreme records (e.g. people diverging from stereotypical representations), improper architectural design in relation to the problem (e.g. use of linear boundaries for complex representations), along with many other reasons.

To be more concrete, consider a potential scenario in credit scoring: an AI system decides whether to discard an application for credit or send it to an analyst for further evaluation. A similar system can be implemented with a family of ML technologies named classifiers. Classifiers are functions which given an input, such as information extracted from a credit application, output a categorical decision within a set of potential choices, for example: $\{0, 1\}$ indicating "*discard*", "*process*". The specific function mapping inputs to output is automatically learned from data providing correct examples of input-output association, through a computational process named *training*. In this context, biases penalizing (or favoring) a certain social category over another (e.g. male vs female) may arise from very subtle statistical differences in data. For instance, if data used for training the classifier contain a percentage of defaults across women significantly superior to men, the algorithm learns to favor men to women in granting credit. This does not imply the algorithm won't grant credit to women, but a specific set of credit applications may exist for which the fact that the applicant is a man or woman determines whether the decision outcome is "*discard*" or "*process*". Notably, in this example the bias (i.e. significantly mean difference in defaults across gender) does not exist in the real world, where women may even default less than men, but appears in data as a consequence of the process of data collections, or merely by chance.

Biases may arise for reasons much more subtle than this. Even if the marginal distributions of defaults are balanced between men and women (i.e. man and women equally default), there may be differences in the joint distributions, which are easily detected by the classifier. This case may also be difficult to diagnose. Another discriminatory behaviour may be generated by the statistical representativeness of groups, is to say, how many observations refer to a group. consider an AI system which is used for sorting applications by creditworthiness: limited resources are available and should be allocated to top n applications. The fact that the algorithm was exposed to a limited number of data for a certain category during training may biases it to assign an inferior score to instances of those category. This behavior reflects the reduced confidence in a group as a result of inferior amount of information.

It should be noted that ML models are designed to detect and take advantage of statistical biases. Indeed, even biases originating from sampling error and raising unfairness are used by algorithms because lead to in-sample performance gain. Some techniques to reduce fairness biases are introduced in Section 6. However, it should be clear that very frequently eliminating a bias implies reducing performance - although the opposite may also be true. Therefore one may wonder why preventing AI from using a discriminatory information (e.g. ethnic group) if this lead to better choices. An answer to a similar question goes beyond the purpose of the present article; however this dilemma is useful to appreciate the intrinsic power of AI technologies in shaping a society were abstract values (e.g. moral values) find concrete application - at least in the digital world.

1.2 Inclusiveness

Inclusiveness requires AI be designed to encourage the widest possible equitable use and access. Inclusiveness becomes crucial when AI is conceived as a resource providing a competitive advantage in social, economic or health domains. The main barriers to inclusiveness are linked to data and computational resources access. From a final user perspective an increasing problem could be having access to appropriate hardware to run AI applications. Today most AI applications run on personal computer, smartphones or are offered as web applications at a reasonable price. In a near future, dedicated hardware, such as GPUs or TPUs are likely to become a standard. Moreover, specific equipment may become necessary to benefit of AI, especially in health-related areas, and equipment may include portable EEG helmets and virtual visors. Being excluded from accessing those devices may become as impacting as being prevented from using a computer today. Another issue is that AI may come in a form not suitable for minorities, because is trained to emulate data which reflect majority group mainly. Furthermore, a benefit of AI is that it may keep improving (i.e. learning) from interactions with final users. However, this pushes AI towards reflecting the majority group.

Inclusiveness also concerns avoiding that some economical entities may benefit of AI far more than others, potentially leading to a new socio-economic equilibrium. Imagine that a certain AI technology allows to predict the stock market with high precision. However, such a technology requires huge amount of data along with unprecedented computational power demand which only a few corporation have or can afford. If this scenario sounds a bit science fiction, it is worth mentioning Generative Pre-trained Transformer 3 (GPT-3). GPT-3 is a language model created by OpenAI, which is capable of generating *ex novo* very realistic human-like text, even concerning technical domains. GPT-3 is considered the most exceptional language model to date. However, GPT-3 does not use any technologies which hasn't already been applied to similar AI architectures, but its higher performance is attributed to its huge number of parameters. GPT-3 capacity is ten times larger than Turing NLG, the next largest NLP model. As a consequence, the amount of data required to train GPT-3 is so huge that the whole Wikipedia represents just 3% of the overall training set. According to one estimate, training GPT-3 costs at least $4.6 million. If a similar technology offers a crucial competitive advantage in any industrial area, only few player would survive.

From a technical perspective, there is little scope for intervention. While there is extensive literature on how to handle AI computational limits, there is no obvious way to extend an algorithm trained with data from population A (e.g. majority group) to make predictions for population B (e.g. minority group). Some recent advancements in ML fields, such as Meta-Learning and Autoencoders, are likely to help developing more generalised forms of AI, which can be extended beyond the data perimeter used for training.

1.3 Reliability and Safety

Reliability concerns refer to AI failures or poor resilience to input changes, and are discussed in greater depth in Section 4. As AI levels off human performance in specific tasks, or even encompasses it, an increasing number of human responsibilities are delegated to AI. However, this implies that a sudden and systematic AI failure is likely to cause serious damage. Once more, this is not a regrettable prevision, but is now a matter of history. The tragedy of lost lives and economic recession caused from the Covid-19 pandemic did not spare AI systems and in particular predictive systems based on ML.

ML forecasts future outcomes based on past events, as described in training data. This fact may suggest that ML cannot forecast an event which has not occurred in training data. Luckily this is not true, otherwise ML forecasting would be reduced to a mere lookup of similar events in past data. However, what ML model in general cannot predict, along with any traditional statistical model, is a change in the logic associating observations to forecasts. Indeed ML models, as most statistical models, are thought to work with Independent and Identically Distributed (i.i.d.) cross-sectional data. Unfortunately many ML applications are not consistent with this assumption. A major violation of the i.i.d. assumption was recently provided by the pandemic. Not only the pandemic damaged AI systems during the time-window including the enforcement of local lockdowns. The influence of the data disturbance caused by the pandemic is still a main issue in any AI application, and data scientists need to incorporate data exogenous to the phenomenon of interest, to factor out the variation caused by Covid-related dynamics. Ordinary phenomena, characterized by little magnitude compared to the pandemic, still have the potential to massively damage the functioning of AI systems.

Reliability concerns are further exacerbated by the fact there is no obvious way to monitor and prevent AI system failures. Although many AI systems provide some kind of confidence point estimates along with predictions, in most cases those confidence should be intended as the probability of the prediction being correct given prediction data is consistent with training data. Therefore, it is not unusual an algorithm make a mistake with high confidence, therefore those confidences usually have little practical utility in monitoring, for example, performance degradation of the system over time. A useful suggestion to improve ML reliability is to investigate prediction errors in training. Those errors reveal far more on ML functioning logic than tons of good predictions. Similarly, it is important to assert that AI properly responds to edge cases, which are typically characterised by a reduce amount of training data (i.e. more uncertainty). The danger is that a very small and apparently insignificant variation in input data may cause an important change in AI behavior. For instance, the Computer Vision system of an autonomous-driving car may exhibit virtually perfect precision in recognizing a stop label as such. However, the same algorithm may be dramatically deceived if someone pastes a yellow smile on the stop label. The amount of red in the label may be a crucial feature for the algorithm to identify the stop label. The small concentration of yellow may push the mathematical representation of the observation in multidimensional regions closer to a "speed limit 130" label, with adverse consequences.

In other terms, changes that do not affect human performance on a task, or which are not even perceivable, may seriously impact AI behavior. As mentioned before, those changes have the potential to make the algorithm very confident in a wrong inference. A proper way to address those issues would require the development of generative models of input data, along with their relation to target behavior. However, this is exactly what ML algorithms attempt to bypass, and also the very reason why ML models found their indisputable success (see Section Alternative Modelling Approaches). If AI systems can be deceived in subtle ways, the answer to reliability issues is to include human workforce in validating the model in a continuous way (Human-in-the-loop). This may lead to the paradoxical conclusion that humans are necessary to supervise systems designed to replace some human activities. We claim that human-AI synergy is indeed the key to successful use of AI in industry - and in particular in finance. Human-AI cooperation is beyond the scope of the present document, but we want to conclude this subsection suggesting that prosperous AI applications should be designed to increase human intelligence, not to replace it.

1.4 Privacy and Security

Privacy is a fundamental right. AI consumes data to learn how to emulate a behavior (training) and to predict. Notably, AI also generates data when doing inference. All these types of data may rise privacy concerns. Input data (either train or test) may contain sensible information which should not be disclosed. A straightforward choice to avoid exposing AI to sensible data is to delete sensible information from input datasets. However, deletion may not be enough, because AI could still recover protected information from the covariance of other variables. Indeed, some algorithms (e.g. DL models) automatically generate internal data representations which are useful to predict the target variables. Those representations may reveal sensible information which are not explicitly contained in either input or output data. For example, a DL model may automatically determine a latent representation which could implicitly infer customers' sexual orientations, religious beliefs and personality profiles. Those representations may help the algorithm recommending products on an online retail platform. However, an expert data scientist can uncover and interpret those representations, and potentially figure out protected information of which the customer himself is not aware. Similarly, an algorithm may be used to directly predict a sensible information.

As any other tools, AI can be leveraged for malicious purposes. In this article we limit to mention some technically interesting AI abuses. In paragraph 2.3 we mentioned as a small change in input data may change the output of an AI system. An expert data scientist could exploit this phenomenon to make subliminal changes to an input in order to obtain a certain prediction outcome, even in absence of appropriate conditions. Noteworthy in this regard is AI be essentially a software, so that all the hacks which can be applied to software (e.g. buffer overflow) could by used against AI systems as well. However, while "traditional" hacks tend to exploit hardware or code logic vulnerabilities, AI is vulnerable in its mathematical aspects as well. Finally, some AI applications can be used to generate data similar to a provided input, or even to generating specific changes with respect to an attribute in inputs. Prominent examples of those applications are fake text, pictures and videos. AI has proved to be able to generate from scratch scientific papers, math textbooks, political speeches; all these examples are formally consistent with originals, but devoid of content. The same applies to media content, where AI can perform classical photo editing tasks and can even generate realistic video depicting people involved in any kind of activity. Clearly, those technologies have an incredible potential to generate ideological and material false representations, with huge social and judicial implications.

1.5 Transparency and Accountability

Accountability refers to the necessity to justify AI predictions to any entities (not necessarily human) interacting with the system. For example, it should be clear whether an algorithm is suggesting to buy a certain product, because customers with similar purchasing history bought it or whether the algorithm is maximising the exposure of paid-advertisement products. Specifically, some families of algorithms (i.e. Reinforcement Learning models) drive the interaction with the final user towards a specific purpose (e.g. maximising sales). Users have the right to know the high-level functioning objective of an AI they interact with, and importantly, have the right not be deceived by AI. This is especially true if AI exploit their data to pursue a goal which is know to the final user. Feedback loops can occur when the model controls the data which appears to the user. The data the user is exposed to can quickly become flowed by the algorithm itself. This typically happen with recommendation system. A consequence is that the user is pushed to consume certain content, as watching a movie or buying a product, but she is deceived to be responsible for those choices. To ensure accountability, predictions should be derivable from the algorithm prediction logic. Therefore such logic should be transparent, although not necessarily deterministic. Transparency includes inspecting model logic and reproducing the dynamics resulting in a certain prediction. Transparency also implies the need for representation of the moral values and societal norms holding in the context of operation, which impact on various step of the algorithm life-cycle, beginning from the very reason which determined its development. Therefore accountability in AI requires both the function of guiding action and the function of explanation.

2. Normative Standard and Guidelines

The regulation of Artificial Intelligence (AI) refers to the development of appropriate policies and legislation aimed at promoting and regulating the adoption of AI technologies. 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. A first line of regulations on data protection, also relevant for AI applications, was enacted by the European Commission under the name of General Data Protection Regulation (GDPR) in 2018. GDPR is related to data which can be linked to identifiable individuals and applies to companies that resides in EU, or provide services in EU, or store data in EU. Notably, fines for violations of GDPR can ammount to 4% of company worldwide revenues. Companies should inform users on the reason why data are collected and can use those data only for the communicated purposes. Moreover, companies should guarantee the security of those data and maintain an active program of monitoring to ensure compliance and accountability of any violations.

In 2019 some AI ethical guidelines were published by the European Commission to manage AI associated risks, while promoting the adoption at the same time. More recently, the European Commission published a proposal for a Regulation on Artificial Intelligence in April 2021. Such a proposal includes several legislative proposal which should be intended as a schema for national regulations, consistent with existing norms. Such European proposal will be summarised later in this section - for a deeper discussion see European Commission Regulation on Artificial Intelligence, iason Research series. Before considering the regulation proposal, it is worth noting that for companies which adopt AI to evolve regulatory frameworks inevitably comes at the cost of reducing ability in the use of AI. On the one hand, with proposals still evolving and a proper legislation still missing, it is not yet clear what should be done. On the other, the process of compliance with the new regulations will take considerable time and may require important technological and methodological upgrades, therefore it should be considered well in advance to avoid a halt of productive systems when a national regulation will enter into force. In the rest of the section the current (proposed) regulatory framework will be discussed from the perspective of the authors - i.e. data scientists technically operating in the field.

Beyond the specific European Commission proposal, we identified three factors which are recurrent in technology normative standards at global level and should be integrated in AI governance practices right now. The first factor is the requirement to conduct assessments of AI risks and to document how such risks have been addressed and minimized, if cannot be solved. This is in line with regulations in force, such as the GDPR, which requires impact assessments for high-risk personal data management. The second factor is accountability and independence, suggesting that data scientists, compliance officers, lawyers and stakeholders evaluating the AI system have different incentives, so that can effectively perform an unbiased evaluation. This may also take the form of hiring outside experts to be involved in the evaluation process to prove full accountability and independence. The third factor is the need for continuous monitoring of AI algorithms, which include the development of appropriate performance metrics. Those metrics can be divided into: outcome metrics, describing the business impact (or the impact on users) of AI solutions, and output metrics, which keep track of the statistical properties of AI outputs, or input-output relations.

Coming to the EU proposal, such *Regulation on Artificial Intelligence* is considered *"the first ever legal framework on AI*â. This document introduces a comprehensive regulatory framework, aiming to provide, from AI developers to final users, with clear requirements and obligations regarding specific uses of AI. The regulation also outlines which technologies shall be considered AI. Those include: any types of Machine Learning (and Deep Learning), Logic- and knowledge-based approaches (including Expert systems), Statistical approaches, Search algorithms and optimization methods. Notably, such definition may be questioned from a technical perspective as these technologies greatly differ one another also in terms of ethical concerns which may rise. Furthermore, the proposal is aware of the efforts needed to change and seeks to reduce administrative and financial burdens for business. The following specific objectives are pursued by the regulation with regard to the use of AI in EU.

- 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.

To achieve these objectives, the proposal adopts a regulatory approach limited to the minimum requirements without i) constraining and hindering technological development and ii) disproportionately increasing the cost of AI development. The European Parliament and the Member States will adopt the Commission proposals in the ordinary legislative procedure, and proposal should be integrated into the existing obligations and procedures under Directive 2013/36/EU (CRD IV). Member States will have to designate one or more national competent authorities (NCAs) for the purpose of supervising the application and implementation of the regulation. The Regulation also provides that some AI practices shall be prohibited. Those include AI systems: using subliminal techniques to drive a person's behavior; exploiting vulnerabilities; evaluating, scoring or classifying the trustworthiness of people (limited to public sector); using real-time remote biometric identification in publicly accessible spaces for the purpose of law enforcement (unless and in as far as such use is strictly necessary).

AI application should be classified based on their inherent risk level. AI systems identified as âhigh-risk AI systemsâ are those used in relation to safety, for instance as safety components on machinery, cableways and medical devices). High-risk AI also includes systems which allow biometric identification and categorization of natural person, those involved in the management of critical operations or infrastructure and those deployed in education or administrative domains. High-risk AI systems are subject to strict obligations. Providers shall implement a risk management system for the entire AI life cycle and ensure that their systems are compliant with the mandatory requirements. Furthermore, providers shall put a quality management system to ensure compliance with the overall regulation, including a technical documentation detailed with information on AI development. For credit institutions, such documentation, along with automatically generated logs, should be part of internal governance - as regulated by Directive (CRD IV) 2013/36 EU, Article 74.

Moreover, before placing on the market or putting into service a high-risk AI system, providers shall register it in the EU database. The registration requires to provide a description of the intended purpose of the AI system and the type and expiry date of the certificate issued by the notified body (along with other information). Certificates are issued by notified bodies after the assessment of conformity with the mandatory requirements. Certificates are valid for the period they indicate, which shall not exceed five years. Where notified bodies finds that an AI system is no longer compliant with the requirements, they shall suspend, withdraw or impose restrictions on the certificate issued. The AI Regulation also imposes obligations on users of high-risk AI systems, who should use AI in accordance with the instructions for use, ensure that input data is relevant, and monitor the operation of the high-risk AI system based on the instructions.

Some obligations are also provided for limited-risk AI system, and mainly concern about usersâ awareness of dealing with AI system. Notably, the Regulation acknowledge that to promote and protect innovation, it is important that the interests of small-scale providers and final users of AI are considered. Finally, the European Parliament and the Member States will need to adopt the Commission's proposals on a European approach for Artificial Intelligence and on Machinery Products in the ordinary legislative procedure and should be integrated even into the existing obligations and procedures under Directive 2013/36/EU (CRD IV).

Before it can be adopted, however, the AI Regulation will join an already crowded digital docket and must pass through a complex and contentious legislative process. The AI Regulation is consistent with the broad outlines of EU policy set out in the Commissionâs February 2020 AI strategy paper, so there are few, if any, complete surprises. However, the broad and potentially vague definitions highlight the difficulty of translating general principles into enforceable legislation. Similarly, the extensive obligations imposed on providers, manufacturers, importers, distributors and users of AI systems will be daunting for all but the largest companies, and the new governance and enforcement regime will add to an increasingly dense regulatory forest in Europe.

3. Technical Considerations on Data-driven Approaches

In the last 15 years there has been considerable advancements in sophisticated AI algorithms which leverage huge amount of data to make decision, forecast, classify or, more generally, optimize a computational problem. Nowadays many AI predictions are comparable to, or even superior, human experts judgment. Not surprisingly, a similar success has generated much enthusiasm around AI applications, leading to the conviction, even among experts, that AI, along with sufficiently big data, allows to make optimal predictions in a purely automated way. In other words, many believe that a sufficiently powerful algorithm could solve any problem (e.g. predict the stock market) without human intervention, as long as provided with a few billion records of labelled data (see later). This section deep dives into such a belief and highlights some issues and limits concerned the automated use of data.

A key feature of AI algorithms is their ability to work out a problem without being explicitly designed for that problem. For example, search algorithms attempt to achieve a certain purpose (e.g. win a chess match) by using heuristics for selecting the best long-term actions (e.g. move), without exploring all possibilities. However, those types of algorithm are not really "intelligent", in the sense cannot handle a certain problem in a condition-specific way, but keep applying the same strategy in any scenario. In contrast Multivariate Pattern Recognition are machines (i.e. algorithms) with the power to *learn*. Arthur Samuel invented some of those algorithms and coined the term âMachine Learningâ in 1952, defining it as a the âfield of study that gives computers the capability to learn without being explicitly programmedâ. Samuel wrote a checkers-playing program which is the worldâs first computer program that can learn from experience. However, it is interesting to note that many Machine Learning techniques have been discovered even before the invention of the first computer. For instance, the method of least squares was published by Legendre in 1805. By 1940s computers appeared, but nobody believe they could learn, until Alan Turing published the legendary paper "Computing Machinery and Intelligence" in 1950. Therefore, Machine learning was not suddenly invented, rather it is an accumulation of techniques from statistics, optimization, algebra, calculus and computer science.

Machine Learning (ML) is the core application of AI, since provides systems the ability to automatically learn from data and improve from experience (i.e. more exposure to data) without being programmed to execute a predetermined strategy. Therefore ML efficiency is proportional to available data, but also to available computational power, necessary for data processing, but also to increase algorithmic complexity. Notably, computing power costs have decreased by a factor of 10 every four years since 1950 - although this trend appears slowing down during the last decade. Similarly, the total amount of data globally generated has grown by a factor of 40 during the last 10 year. Therefore, it is not surprising that while quantitative domains such as risk management, decision making and finance were traditionally led by theory-related models (theory driven approach), during the last 15 years the trend reversed with the increasingly adoption of powerful algorithms for pattern recognition and automated prediction (data-driven approach).

Although, data-driven approaches provide many important advantages, it is reasonable to expect that high-risk AI applications be designed in a thoughtful manner, by taking into account knowledge on the process being modelled and not just data. On the contrary, data-driven approaches are rather problem-agnostic and learn to emulate a behavior without necessarily grasping the underlying dynamics. This has led to draw more attention on "brute-forcing" problems, for example by enlarging model computational complexity, instead of increasing the comprehension of the problem itself. More recently, the widespread adoption of data-driven algorithms generated concerns, mainly regarding two crucial and related points: how much can we trust them (uncertainty about prediction) and how much control can we have on them (uncertainty about predictive logic). Those concerns are also motivated by empirical evidence showing that, although ML has proved excellent performance peaks, it is vulnerable to subtle changes in input data (i.e. small changes in input cause dramatic changes in output). Furthermore, ML is vulnerable in an unpredictable way, because leverages decision-criteria which do not resemble human reasoning. This also implies that the decision-criteria allowing the algorithm to work are usually very unclear. Finally, it should be noted that complex tasks, such as trading the financial market, are not executed by a single algorithm, but by a wealth of algorithms which interact one another, generating further complexity and unpredictability.

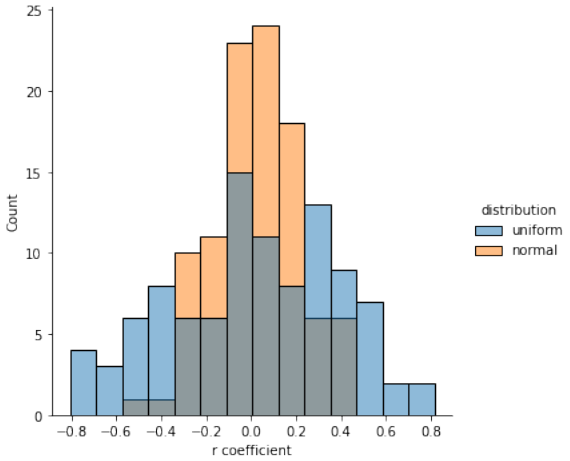


FIGURE 1: Person r distribution for independent samples drawn from Normal or Uniform

As previously mentioned, the term *AI* encompasses several technologies which are combined to emulate a certain behavior. Therefore an AI architecture is typically composed by different modules, each tackling a specific logical or computational problem, overall exhibiting an âintelligentâ behavior. Such a behavior is not limited to return a prediction, but also includes the dynamic management of a wealth of processes taking place underneath the hood, as for example converting a web-page into a list of tokenized sentences, and in turn those sentences in mathematical representations which can be handled by statistical models. Consider an AI architecture which analyses balance sheets, financial news bulletins, historical stock prices (along with other relevant information) to forecast the closing price of a stock. Within such an architecture some algorithms operate in a deterministic way, as are explicitly programmed to carry out specific tasks (e.g. text extraction and normalization). Notably, several issues, including fairness and reliability concerns, may already arise at this stage. Other algorithms use statistical or mathematical models of varying complexity. Those models are characterised by assumptions, that should be met to guarantee the validity of results.

For example, the ubiquitous use of Pearson correlation coefficient (r) should be considered with respect to its assumptions, namely: continuous variables, related pairs, absence of outliers, variables are normally distributed, variables are linearity related and homoscedasticity (stable variance) holds. Figure 1 illustrates the result of correlating 100 independent pairs of random variables drawn from a Normal or a Uniform distribution. Notably, in the latter case the normality assumption is violated. In either case, the correlation coefficient is not quite 0 for a certain portion of samples. However, while independent samples drawn from a Normal distribution have on average a Pearson coefficient between $[-0.2, 0.2]$, if the normality assumption is violated the probability of observing $|r| > 0.7$ with independent variables is greater than 5%. Therefore, violating model assumptions leads to decreased predictive performance, increased uncertainty and explainability issues.

Finally, other algorithms operate in an almost impenetrable manner (*black box*), and there is no clue how inputs are connected to outputs. This is typically the case for Deep Learning models - although important difference exists. For most of those algorithms, only the contribute of each variable (i.e. feature) with respect to the prediction can be somehow estimated, for example by comparing what a model predicts with and without that attribute. However, complex models are far from using the linear contribution of each features, therefore a very small change in inputs may correspond to a completely different output. For instance, with reference to the trading AI system, it may be reasonable to guess that if there is no relevant news concerning a certain stock in a specific day, any neutral change to the text should not affect the final prediction, as the text does not contain any useful information in relation to the stock price. However, this belief is likely to be contradicted. A similar phenomenon may suggest a structural flaw in the architecture, but indeed it simply reflects the fact the algorithms does not âreasonâ according to logic or causal relations. Instead, ML projects the input information (i.e. features) into a complex mathematical space, which is segmented into portions associated to outcomes. Depending on the algorithm, such projections may be recovered or be extremely hard to decode.

One may wonder whether such a level of complexity is really necessary. Surprisingly, most often the answer is no. Model complexity should be proportional to the intrinsic complexity of the phenomenon being modelled. It should be noted here that AI systems emulating cognitive abilities, such as image recognition or natural language understanding, are indeed approximating the functioning of far more complex systems - namely the neurophisiologycal networks in the brain. In contrast, most AI applications in industry do not need to resemble a complex dynamical system composed by 100 billion interacting elements. If the association between input data and outcomes is not too abstract, simpler models may outperform sophisticated Deep Learning algorithms. However, what frequently happens is a data scientist be asked to model a phenomenon he is totally unaware of. The fast and easy solution becomes to throw a bunch of data to some data-driven pipeline and increase ML complexity hoping in improving the prediction outcome. Complexity could be decreased and predictions improved if some hints are gained on the data-outcome relationship and such knowledge is encoded in the feature engineering process.

From a mathematical perspective, optimal inference on y based on observed events X can be obtained without any AI architectures by computing $P(y|X)$. However, such a probability distribution cannot be computed directly for most use cases, because of both analytical intractability and information limits on $P(X)$. Therefore, the power of AI lays in several families of computational models which leverage mathematical tricks to approximate $P(y|X)$ as a function $f(X, w) \rightarrow y$, where w is a set of parameters determining the function behavior. Those computational models are the ML (and DL) algorithms discussed so far. ML models are trained on data, resulting in deterministic algorithms which map input to output using a point estimate of parameter weights, usually optimized by Maximum-Likelihood. From this point of view, a major concern is that ML algorithms are optimized on finite data subsets coming from an unknown distribution, therefore different models and different parameter weights may fit the same data equally well, yet none of them being accurate with respect to the population from which data are drawn. Considering that the population is usually unknown, the only solution to mitigate this issue is to select models and weights consistent with the underlying phenomenon originating the data (see Section Empirical Fairness in AI).

3.1 ML Performance Degradation Under Distributional Shift

A very simple experiment was set up to illustrate the effect of a distributional shift. A list of height and weight pairs (N=25k) was considered to train a ML regressor to predict height (y) from weight (X). Figure 2 shows how height linearly grows with weight for most of the distribution around the mean, while tails tend to deviate. Such a pattern can be observed in many practical phenomena.

In order to simulate a shift between train and prediction distributions, observations were assigned to two groups (group 0 and group 1) with an aleatory function which increases the probability of assigning an observation to group 1 as weight increases. As a results, the average weight is slightly lower for group 0 than group 1; considering the observed linear relation, the same applies to height (Figure 3). Notice that a shift in distributions means does not represent an issue per se, indeed an adequate modelling procedure is expected to generate a model which can generalize to unseen

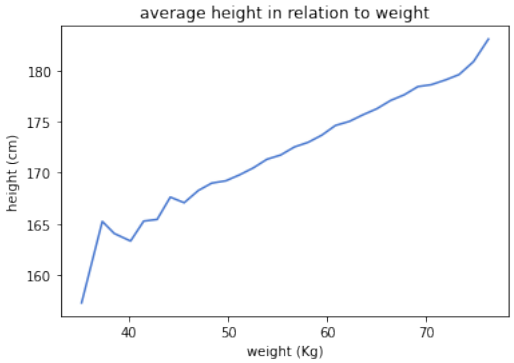


FIGURE 2: Plot of height in function of weight

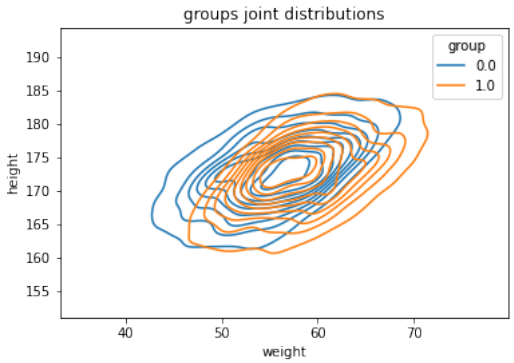


FIGURE 3: Plot of the distributions of the two groups (0 and 1) used for training and test purposes

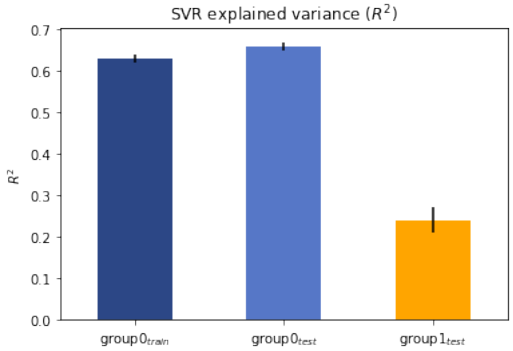


FIGURE 4: Plot of SVR performance in terms of R^2 (explained variance)

scenarios, at least to a certain extent. However it is fundamental the shift in means does not implies a change in the relation between X and y - as partially occurs in this example.

A Support Vector Machine Regressor (SVR). as implemented in scikit-learn, was used for modelling. A 3rd-degree polynomial kernel was used and relevant hyperparameters were estimated from trainset noise. Group 0 data were split into train ($group0_{train}$) and test ($group0_{test}$) independent samples for cross-validation, while Group 1 data were entirely used for testing $group0_{test}$. Therefore SVR was trained on $group0_{train}$, then tested on both $group0_{test}$ and $group1_{test}$. In this simulation Group 0 represents the data available during AI development, while Group 1 simulates observation occurring when the algorithm is deployed in production. Figure 4 summarises the results in terms of explained variance (R^2), a measure of how much variation in the target variable (i.e. height) is captured by the model (where $R^2 = 1$ means perfect predictions). While $group0_{train}$ and $group0_{test}$ show appreciable and similar R^2, indicating a good fit of the model and absence of overfitting effects, R^2 appears reduced by half in $group1_{test}$.

Several factors contribute to the R^2 drop in $group1_{test}$. It is worth mentioning how the curve of height in function of weight changes for extreme height values in a way which cannot be predicted from Group 0 data. This is probably not the main source of fall in performance in this case, but highlights an important concern, which it may be described as "looking back when one should look forward". To describe such a problem an example is introduced: suppose we want to predict the surface area of a balloon after blowing n times into it. A ML algorithm would be very successful in predicting the surface area resulting after n blows, potentially modelling subject differences in the effect of blowing, possible cycles, the non-linearity of the surface increase due to the physic of the phenomenon. The only thing the ML algorithm cannot predict is the most obvious: the balloons burst at a certain threshold. Standard ML models cannot incorporate this information unless several balloons bursts are reported in training data - but bursts are what typically we want to prevent.

4. Alternative Modelling Approaches

Popular ML libraries, such as Scikit-learn, provide hundreds of ML models implementations available for free, tweaked with recent innovations from ML papers and ready to use with high-level APIs. Most of those ML models work just fine out of the box and provide a ratio between performance and development costs which is hard to beat. Similar pre-trained DL implementations are freely offered by several repositories, and include state-of-the-art transformer architectures, such as a limited version of GPT-2, which would be practically impossible to train from scratch for most business uses. Comprehensibly, ML represents a targeted solution for many use cases. Some Cloud providers offer services which automatically select and train a ML solution based on uploaded training data. All of this was hardly conceivable just ten years ago and it is certainly a great benefit for several economic sectors. However, there are high-risk domain where ML, while still useful, is not enough in terms of either performance (including reliability) or fairness.

First of all, some domains are characterised by hundreds of years of research and it would be reasonable to require an algorithm to reflect some of this knowledge, instead of learning from scratch from limited data. Interestingly, some fields of study are characterized by "noisy data", violating the ML requirement of cross-sectional i.i.d. data. Common examples of "noisy data" in financial applications are censored/truncated data and data selection biases. In the first case the dependent variable is observed (or recorded) only within a certain interval and exceeding values are not recorded at all (censored) or truncated to interval boundaries. In the second case, available data are not representative of the population they are drawn from, for example because records are collected at the end of a selection process (e.g. application approval). In both cases, fitting any ML models on those data will result in biased (i.e. systematically wrong) predictions. Finally, in some applications knowing the range of variation of a prediction is more important than knowing the most likely occurrence. For instance, we may be more interested in knowing the probability a certain index drops below a certain critical threshold, than predicting its most likely value.

Probabilistic methods allow to provide an answer to the forementioned problems. On the one hand, probabilistic methods allow to encode prior information into the inferential process; for example, information about data being censored/truncated or sampled with a bias. On the other hand, probabilistic methods offer the possibility to switch mindset about the very purpose of estimation. In fact, from a Bayesian perspective, there is no true parameter to approximate with maximum precision, as unknown parameters are indeed random variables. This means that the variation empirically observed, as the difference in samples means drawn from the same population, are not only attributable to sampling or measurement error, but reflect an intrinsic variation of the population parameter itself. Random variables can be described in advanced, before incorporating data, to encode knowledge on a phenomenon into the model estimation. This mindset allows to compute the parameters posterior distribution, and optimize with respect to the whole distribution of what a parameter value could be. In other terms, while classical approaches uses Maximum-Likelihood estimation to find the best parameters for the data under analysis, Bayesian approaches find a distribution of parameters for all possible data occurrences.

4.1 Probabilistic Methods

Probabilistic models rely on Bayesâ theorem to incrementally integrate evidence provided by each variable into a generative model of the phenomenon under study. Although probabilistic model may appear completely unrelated to classical ML, these two families of inferential models are conceptually comparable and can be combined in a unified architecture. A *Perceptron*, the simplest form of Neural Network (more complex forms give rise to Deep Learning), can be seen as a graph where each input variable is connected to all the other variables through a set of randomly initialized weights, which should converge to some optimal value during training. In contrast, a probabilistic model is a graph where a variable has some type of connection only to variables related to it via a conditional probability distribution, so that $A \rightarrow B$ becomes $P(B|A)$. The type of connections, the connected variables and the resulting probability distributions are initialized by a data scientist and should reflect knowledge about the underlying dynamics. The training process optimizes likelihood parameters (i.e. observed variables) through Maximum Likelihood, as in classical ML. However, latent variables distribution (corresponding to weights in ML) are updated accordingly to Bayes'

rule. This process results in a set of posterior distributions describing the probability distribution of all the possible variations. An exhaustive description of probabilistic methods is far beyond the purposes of this document, but some key features are mentioned and briefly discussed.

A probabilistic model allows to specify a graph similarly to a Bayesian network, but it can leverage computational tools, such as $if - else$ statements, which cannot be incorporated in a Bayesian network. When data enter into the graph, a propagation mechanism updates probability distributions in the network, in any directions. Therefore, probabilistic models allow a type of backward inference named *explaining away* (nonmonotonic reasoning), which enables inference not only on the target variables, but also on all the other observed or unobserved (i.e. latent) variables in the network. So entering an observation in a (effect) node will result in back propagation, meaning updating the probability distribution of a cause node (and vice versa). Notably, those networks do not require a complete set of observations (i.e. all the features) to make predictions, as needed in ML. Indeed, the model can even work without any observation at all, in such a case it would assume the prior distribution. Finally, a probabilistic model is transparent in the way it arrives at a decision, which is very beneficial for managing the fairness issues discussed before.

Probabilistic models are trained through Monte Carlo Markov Chain (MCMC) or Variational Inference (VI). We leave the details of those techniques, but we limit to say that MCMC attempts to approximate the posterior distributions by sampling from it and moving towards region characterized by increased probability densities. This procedure is usually slow and may require some tricks to work properly, but potentially can provide very good estimates for any type of graph. In contrast VI approximates the posterior by fitting a family of arbitrary distributions (guide) through Maximum Likelihood. Therefore VI is an optimization procedure, technically very similar to Stochastic Gradient Descent used in Deep Learning. VI is fast and easy to parallelize, but its performance greatly depends on the choice of the guide, which is not an obvious choice.

4.2 Beyond Predictions

In the present document we have discussed several concerns arising from algorithms which make predictions. Those predictions may be unreliable, biased and obtained in a black-box manner. ML is useful to make predictions when the dynamic generating the outcomes is unknown or cannot be formally described. But ML has little to enrich knowledge on the problem is applied to. In a previous section, GPT-3 was mentioned, a colossal language model capable of generating human-like text. What did we learn about language from GTP-3? Nothing. We argue that in many scenarios even more relevant than predictions is the ability to identify and apply knowledge. In this context, knowledge can be conceived as a set of correlative, causal and logical relationships, which organises the information reflected in data. Such an organization principle can be modelled by a probabilistic graph, bringing several benefits.

As ML models, probabilistic graphs can be used to predict and make decisions. But they can also be used to design, communicate and, notably, understand observable dynamics. A great advantage of graphs resides in their potential to reveal conditions under which results hold, or revealing which conditions would lead to different results. This allows to generate counterfactual observations, or more in general operate counterfactual reasoning. For instance, a common AI application in financial industry is the estimation of creditworthiness. Such a task poses an important problem. Typically, banks have records for clients to whom credit was granted. The vast majority of those clients repayed the debt, leading to a severe representativeness unbalance between good and bad debtors. Therefore, a similar dataset, just as it is, cannot be used for training an algorithm to score creditworthiness, as there is not enough information to model the distributions of clients who defaulted (little information) and potential good borrower who haven't got credit (no information at all). However, Bayesian networks can handle this type of problems. Those stochastic networks can generate all possible scenarios conditioned to a set of observations, therefore can emulate the outcome of lending money to a client who had his credit application rejected (or did not apply at all). Importantly, the result of the simulation is not just a single prediction but a whole probability distribution. This is crucial to make decisions that take into account the risk level. Furthermore, this type of models allow to determine which attribute would lead to a different outcome. In theory, these techniques can be used to direct the client to actions oriented to the desired outcome (e.g. obtain credit), or to reach an optimal agreement which is beneficial to both parties.

5. Empirical Fairness in AI

In the previous section computational approaches alternative to ML were discussed. Those approaches can help reducing and managing fairness issues. However, whether we want to avoid, detect or fix fairness-related problems in AI, an operative definition of those kind of problems is needed. Indeed, the pursuit of fairness in AI frameworks poses a further issue, which hasn't been mentioned yet. As AI is essentially composed by mathematical and computational methods, a descriptions of "fairness" in a formal language - such as mathematics or computer code - is necessary to tackle related ethical issues. Considering that common definitions of ethics are:

- "The discipline of dealing with what is good or bad" (Websterâs Ninth New Collegiate Dictionary);
- "The branch of philosophy that deals with distinctions between right and wrong" (J.M. Last);
- "A set of moral principles, especially ones relating to or affirming a specified group, field, or form of conduct" (Oxford Dictionary).

It is hardly surprising that operational definitions of ethical concepts, such as fairness, are somehow shallow, deficient and, notably, conflicting one another. Therefore, there is no standard way to assess the fairness of an AI system and data scientists performing fairness evaluations are required to understand which practices best apply to each scenario. In the following subparagraphs the problem of fairness in AI is broken into semi-independent components on which can arise and can be tackled.

5.1 Metrics Concerning Fairness among Groups

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:

1. **Demographic Parity**: suggests that a predictor is unbiased if the prediction $\hat{y}$ is independent of the protected attribute A, so that:

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

   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.

2. **Equality of Odds** [6]: a predictor $\hat{y}$ satisfies *equalized odds* with respect to protected attribute A and outcome Y, if $\hat{y}$ and A are conditionally independent on Y.

$$\begin{aligned} &Pr\{\hat{Y} = 1|A = 0, Y = y\} \\ &= Pr\{\hat{Y} = 1|A = 1, Y = y\}, \quad y \in \{0, 1\}. \end{aligned}$$

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

3. **Equality of Opportunity** [6]: a binary predictor $\hat{y}$ satisfies equal opportunity with respect to A and Y if:

$$\begin{aligned} &Pr\{\hat{Y} = 1|A = 0, Y = 1\} \\ &= Pr\{\hat{Y} = 1|A = 1, Y = 1\}. \end{aligned}$$

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

4. **Predictive Parity [8]**: 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).

5.2 Bias Mitigation Algorithms

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 (by Change thresholds or Trade off accuracy for fairness). Methods to address biases are discussed for each of those stages.

5.3 Data Collection

Identify biases intrinsic to data is fundamental, because those not only cause biased predictions for protected groups, but may invalidate the entire decision function estimated by the algorithm. Biases may arise in data in several forms, such as mean or variance differences among groups and covariates. Interestingly, even when biases do not appear in data, still data may rise biases in the algorithm predictions. This is typically caused by differences in classes representativity - along with more complex cases. There is no obvious method to address those issue, but collect more data. However, this is typically hard to do, because data collection may be expensive and tend to replicate the same biases, if not performed according to statistical sampling strategy (which are even more expensive).

5.3.1 Pre-Processing

A 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.

Four approaches for removing bias by manipulating the dataset are now described.

- **Manipulating labels [3]**: 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.

- **Manipulating observed data [5]**: proposed manipulating individual features x from data X in a way that depends on the protected attribute p. Each dimension x was divided in case where the protected attribute p is 0 or 1. Then the cumulative distributions $F_0[x]$ and $F_1[x]$ are computed and aligned to a median cumulative distribution $F_m[x]$. Conceptually, this procedure is similar to standardising group of observations by subtracting the mean of each group. This approach has the disadvantage that it treats each input variable $x \in X$ separately, which may affect the joint distribution of X (e.g. feature interactions).

- **Manipulating labels and data [4]**: learned a randomized transformation $P(x', y'|x, y, p)$ that 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 formulate 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 taking into account interactions between data dimensions. However, the randomized transformation is formulated as a probability table, so this is only suitable for datasets with small numbers of discrete input and output variables.

- **Re-weighting data pairs [3]**: 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.

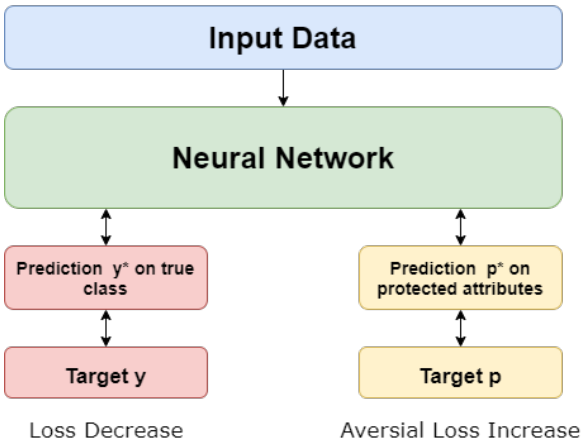


FIGURE 5: Adversarial learning for fairness. While a first algorithm learns a decision function, a second model exploits such function to predict the p. This techniques attempts to maximize predictions while minimizing p effect on predictions

The mitigation techniques applied during preprocessing increase dataset quality but can only be used to optimize Statistical Parity or Individual Fairness (if the metric is given) because they do not have any information on classification labels. Furthermore, these actions will decrease performances (mainly accuracy) with respect to other methods applied outside preprocessing.

5.3.2 In-Processing

An elegant (but complex) approach for removing bias during training is to explicitly remove such a dependency by leveraging Adversarial Learning. Other easier approaches include penalising the Mutual Information between p and the prediction using regularization techniques and fitting the model under the constraint that it is not biased. All those methods are briefly discussed.

- **Adversarial de-biasing [2, 7]**: 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 5 illustrates the Adversarial de-biasing mechanism.

- **Prejudice removal by regularization [1]**: proposed adding an extra regularization condition to a classifier for minimizing the Mutual Information between the protected attribute p and the prediction $\hat{y}$.

- **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.

- **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.

5.3.3 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 $\hat{y}$ to enforce a specified fairness constraint. Therefore post-processing approaches provide great flexibility and do not require model retraining.

6. Conclusions

AI plays an increasingly important role in finance, and today is hard to find a use case in stock price prediction or credit ratings where ML is not being leveraged in some form. However, AI systems do not come without risks and remarkable ethical implications. A crucial issue is that most AI applications work in a black-box manner, behaving according to unintelligible mathematical logics which do not resemble human reasoning (or formal logic). Moreover, most ML algorithms were originally designed to work on cross-sectional data and are sensible to structural changes induced by time.

Notably, ML limits may be more relevant to finance than other high-risk industries, as gains in this domain are frequently connected to benefit from systematic mistakes of another player. In this context, we do not believe that a particular AI technology by itself can offer an important competitive advantage, while a systematic weakness in a deployed AI solution is likely to draw significant disadvantages. We claim knowledge make the difference, and ML should be used beyond mere predictions, to test and inform theories that explain data. Only theories can individuate cause-effect mechanisms that allow to make reliable predictions and decision, which are not implicitly contained in data. A theory can explain factual evidence as well as counterfactual cases. A theory provides knowledge which, ultimately, is what allows to profit on unprecedented circumstances, where the crowd (of ML models) fails.

In this document we introduced probabilistic models, which carry multiple advantages in comparison to ML. They are flexible enough to encode high-level knowledge into the algorithm and are also transparent in the way elaborate evidence to reach a conclusion. These two elements are promising to solve many (but not all) problems concerning AI fairness. Furthermore, probabilistic models allow to increase knowledge on the phenomenon being investigated, which can lead to a virtuous cycle of increasing productivity and predictive performance. However, probabilistic models are hard to implement, depending upon very advanced statistical, computational and domain-specific competences, but also prolonged computations time. Overall, they require increased investments in terms of expertise, development times and costs.

References

[1] **Akaho S., Kamishima T. and Sakuma J.** *Fairness-aware learning through regularization approach*. IEEE, January 2012.

[2] **Beutel A., Chen J., Chi E. and Zhao Z.** *Data decisions and theoretical implications when adversarially learning fair representations*. ArXiv, July 2017.

[3] **Calders T. and Kamiran F.** *Data preprocessing techniques for classification without discrimination*. Springer, December 2011.

[4] **Calmon F., Ramamurthy K., Varshney K., Vinzamuri B. and Wei D.** *Optimized pre-processing for discrimination prevention*. NeurIPS, 2017.

[5] **Feldman M., Friedler S, Moeller J., Scheidegger C. and Venkatasubramanian S.** *Certifying and removing disparate impact*. arXiv, July 2015.

[6] **Hardt M., Price E. and Srebro N.** *Quality of opportunity in supervised learning*. NeurIPS , 2016.

[7] **Lemoine B., Mitchell M. and Zhang B.** *Mitigating unwanted biases with adversarial learning*. ArXiv, January 2018.

[8] **Rubin J. and Verma S.** *Fairness definitions explained*. IEEE, May 2018.

iason is an international firm that consults Financial Institutions on Risk Management. iason is a leader in quantitative analysis and advanced risk methodology, offering a unique mix of know-how and expertise on the pricing of complex financial products and the management of financial, credit and liquidity risks. In addition iason provides a suite of essential solutions to meet the fundamental needs of Financial Institutions.

Techneshtai is a technology start-up specialising in providing Tech solutions to the financial sector.
Technesthai supports companies to improve their businesses with bespoken technology ecosystems: the solutions range from Artificial Intelligence and Machine Learning researches to the implementation of custom web-based applications.
Techneshtai is the ideal technology partner to meet the major challenges that digitisation is imposing on the financial sector.

iason
ESSENTIAL SERVICES FOR
FINANCIAL INSTITUTIONS

TECHNESTHAI