A SYMBOLIC APPROACH FOR ENSURING FAIRNESS IN AI

The design of an explainable AI

Marco Billi

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista semestrale on-line: www.i-lex.it
Luglio 2021
Fascicolo 1
ISSN 1825-1927

i-lex. Scienze Giuridiche, Scienze Cognitive e Intelligenza Artificiale
Rivista semestrale on-line: www.i-lex.it

A SYMBOLIC APPROACH FOR ENSURING FAIRNESS IN AI
THE DESIGN OF AN EXPLAINABLE AI

MARCO BILLI[*]

Abstract.
In the wake of the fourth industrial revolution, the range for potential uses of AI has increased, and with it its potential for harm. Nowadays, new applications are being consistently developed in the fields of healthcare, climate change regulation, security, workplace management and control, and governance mechanisms. At the same time, AI entails a number of potential risks, from being wilfully used for criminal purposes to unexpected and potentially harmful consequences for the health and freedom of its users, such as gender or race based discrimination, transparency and privacy concerns, and opaque decision-making. In following this trend, EU institutions have given rise to a prolific regulatory framework of principle-based ethics codes and guidelines. The scope of this paper is to analyse from a European legal perspective the current fairness principles regulating the design of AI systems, the relevant legal concerns in case law, and how a symbolic approach to AI could help insure fairness, interpretability and transparency. In particular, the focus shall be on the latest projects concerning computable representation of EU directives and how to build upon those to augment the efficiency of such systems and their adherence to the aforementioned ethical principles.

Keywords: *Symbolic, fairness, explainability*

1 Introduction

In the wake of the fourth industrial revolution, the range for potential positive uses of AI has increased, as has its potential for harm. In order to offset such consequences and counteract the opacity of AI

[**]Alma AI – Alma Mater Research Institute for Human-Centered Artificial Intelligence, Alma Mater Studiorum—Università di Bologna, Italy, marco.billi3@unibo.it (0000-0002-6807-073X).

systems, the EU Commission has set out a strategy for the development of 'trustworthy AI'[1].

In April 2018 the Commission established a High-Level Expert Group that, after a year of open consultation and comments, published “Ethics guidelines for trustworthy AI”, a document citing seven key requirements for AI systems to be considered trustworthy[2]. These requirements were adapted into principles by the European Commission for the Efficiency of Justice,[3] to ensure compliance in the processing of judicial decisions and data by algorithms and the use that is made of that data.

With regard to data itself, good management is a key conduit to research and innovation. Data is an integral component of both human and machine driven activities, thus a specific set of guidelines for its collection and use was the subject of another Expert Group that led, between 2016 and 2018, to the development of the notion of “fair” data[4]. Furthermore, the GDPR has set specific requirements for its access and scope, and personal data is subject to several additional limitations.

The principle of fairness is highlighted in the aforementioned legal sources, establishing its importance in the European legal system This leads to the question of, how this principle should be translated into practice, in the absence of any standardized process.

The first chapter of this paper shall introduce the concept of fairness of AI algorithms, a growing field of research that arises from the

[1] The very first draft required the respect of fundamental rights and the robustness and reliability of a system to be considered ethically compliant and trustworthy, the copy can still be found at https://www.euractiv.com/wp-content/uploads/sites/2/2018/12/AIHLEGDraftAIEthicsGuidelinespdf.pdf

[2] The Guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy: 1) human agency and oversight; 2) technical Robustness and safety; 3) privacy and data governance; 4) transparency; 5) diversity, non-discrimination and fairness; 6) societal and environmental well-being; 7) accountability.

[3] The CEPEJ has identified the following core principles to be respected in the field of AI and justice: 1) Principle of respect of fundamental rights; 2) Principle of non-discrimination 3) Principle of quality and security 4) Principle of transparency, impartiality and fairness 5) Principle “under user control”.

[4] First published in the Scientific data journal https://www.go-fair.org/fair-principles/

general need for decisions to be free from bias and discrimination. The more common ways it can be implemented in the engineering process will then be discussed, as will the implications of the EU regulatory framework, including the White Paper on AI, the CEPEJ Guidelines and the FAIR data principles.

Single Member States such as Germany or Denmark have also started to establish their own internal accountability standards, focusing on the creation of certification systems or licenses to ensure the respect of principles of non-discrimination and fairness.

The German Data Ethics Commission[5], for example, recommends adopting a risk-based approach to algorithmic systems, which has been at the forefront of the latest regulation proposal of the European Commission. The principle underlying this approach is as follows: the greater the potential for harm, the more stringent the requirements and the more far-reaching the intervention by means of regulatory instruments.

Especially in the domain of law-making and dispensation of justice , the German commission states that algorithmic systems “may at most be used for peripheral tasks ”, and ”not be used to undermine the functional independence of the courts or the democratic process”, although an exception should be made for all administrative tasks relating to the provision of benefits and services. Furthermore, it remarks that decisions made by the State on the basis of algorithmic systems must still be transparent, and it must be possible to provide justifications for them.

The burden of providing transparency and explainability of AI decisions is left to the programmers and engineers, without clearly defining how and why.

Certain companies have been trying to push for the creation of an internal auditing mechanism to be followed in the engineering process involved in AI creation, such as Google or Microsoft.

[5] The last meeting was held in August 2019
https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html

Selbst and Barocas [6]argue that “one must seek explanations of the process behind a model’s development, not just explanations of the model itself”.

This technocratic interpretation of the ethical principles shifts the focus on the development process, stating that, on the one hand, as long as the safety procedure has been followed the system should be made available to the public without further concerns.

This point of view, on the other hand, can hardly be called satisfactory. Even if governments, following this trend, gave a seal of approval to all AI systems that comply to the mechanism, the legal domain, and especially the automated decision-making area of research, would be woefully unprepared for an influx of algorithmic machines that pose themselves as fully compliant to the fairness principle without any possibility of acquiring evidence to the contrary.

Without even taking into account the current GDPR regulation that, although with some exceptions, prohibits fully-automated decision-making AI, once two systems give diametrically opposed answers, without a solid explanation of the reasoning process it would create needless confusion and uncertainty for any user.

While this approach may work for certain tasks and services, domains such as law and healthcare will need more stringent requirements.

In the second chapter of this paper we shall take a look at how the use of such quantities of data, accordingly named big data, is particularly susceptible to result in the violation of basic rights, as bias in the training data can be hard to spot before the final result is already in deployment.

The need to focus on understanding the reasoning process of an AI system has seen its rise ever since the advent of consequential mechanisms of classification and ranking, such as spam filters, credit card fraud detection, search engines, tailored news trends, market segmentation and advertising, insurance or loan qualification, and credit scoring. These are just some examples of classification systems that gather personal data and make an impact in the daily life, in our network-connected, advanced capitalist societies.

[6] Barocas, Solon and Selbst, Andrew D., Big Data's Disparate Impact (2016). 104 California Law Review 671 (2016), Available at SSRN: https://ssrn.com/abstract=2477899 or http://dx.doi.org/10.2139/ssrn.2477899

These mechanisms frequently rely on computational algorithms, and on machine learning algorithms to do this work.

Opaqueness is now an important concern among legal scholars and scientists. The algorithms in question operate using vast amounts of data as input, through which they produce an output: for example, a classification (i.e. whether to give an applicant a loan, or whether to tag an email as spam). It is often the case that the recipient of a classification result is not made aware of the reasoning that brought to that conclusion, therefore the system is considered opaque.

The final chapters will approach the fairness principle from a practical perspective, using a symbolic approach, focused on an isomorphic representation of the original legal source, which can be used to assess the interpretability and explainability of an AI.

In contrast with ML systems, symbolic applications process strings of characters that represent real-world entities or concepts. Symbols can be then arranged in structures such as lists, hierarchies, or networks and these structures show how they relate to each other.

The main advantage of symbolic programming for AI is the ease of creating and manipulating complex data structures, which in our case reflect the complexity of the law, and it facilitates the writing of domain specific languages. A computer program is finally able to make sense of the interactions between those symbols to produce the expected result, and the logical inferences used can later be translated in natural language.

The scope of the EU project CrossJustice is to represent the rights of defendants in criminal matters in the different Member States in a computable standard accessible by professionals and citizens alike.[7]

The goal is not only to help legal practitioners in their daily activities, but to build a platform that helps interoperability and communication between the several regulations that single member states have adopted, how these regulations interact, and their closeness to the EU directives.

The main advantage of using symbolic representation lies in its easily explainable logical process, developed by human programmers and subsequent transparent result.

[7] The CrossJustice project is co-funded by the Justice Programme of the European Union (2014-2020).

Lately, although symbolic approaches have fallen out of fashion, neuro-symbolic ones have started to surface, providing a solution to the black box (opaqueness) problem. Merging the efficiency and scope of ML methods and the explainability of symbolic machines will be the key in future undertakings.

2 The Fairness principle

The issues that Courts are now facing are two major, connected, ones. The first refers to the possibility that algorithmic decision-making may lead to discriminatory decision, either due to a defect in the structure of the program or to bias in the data embedded in the training process. The second refers to the lack of information that is often found in AI systems, which cannot be properly explained in human-comprehensible terms, regarding the logic of their decision-making.

The meaning and purpose of the fairness principle, which aims at alleviating the two issues, encompasses several legal aspects.

When the EU[8] first introduced it, the text made express reference to the principles of “diversity, non-discrimination and fairness”, and its solution was based on ensuring the participation of multiple actors in the engineering stage, so that the final result would hopefully be free from any bias in the data used for the training of the system.

All algorithms are developed based on previously collected data, be it legal sources, machine-made correlations, or real-world instances. If the data collected is biased or lacking, it will be reflected in the system. The aim of the trustworthy AI principles is thus rooted in making sure that such data will be comprehensive enough to ensure an impartial and just treatment in its use.

The problem of the opaqueness of a system is partially covered by the principle of transparency, which acts as separate from that of fairness, and introduced the notion of “traceability” of AI systems[9]. It is stated that a mechanism to log and document both the decisions taken by the system, and the path it used to reach that decision, should be ensured. The principle of transparency goes together with explainability, which and aims at providing explanations to the user of a system of both the decision, and the degree to which an AI system influences

[8] Directive Proposal (COM(2019)168).

[9] EU Commission, Ethics Guidelines for Trustworthy AI, 2019.

and shapes the organizational decision-making process, design choices, and rationale behind the system itself.

It should be noted that explainability and interpretability are often used interchangeably. While the former refers to the understanding of both what a node represents in the system and its importance to the model's performance, the latter should be construed as the ability to determine cause and effect from a model.

The White Paper on AI makes express reference to this distinction as dictated by the COM, while the CEPEJ, in its guidelines takes a slightly different approach.

As far as the principle of non-discrimination is concerned, the Council of Europe does not concern itself with the parties involved in the development process, but relies instead on limiting and neutralizing existing discriminatory practices and giving careful consideration to the sensitive data that can be processed directly or indirectly when training the algorithm. Basically, it gives priority to the existing data, ensuring that no bias is introduced to the development and deployment process, in contrast with the previous objective, the collection of vast amounts of diverse data.

Even more interesting is the fourth principle: “transparency, impartiality and fairness”. Here the EU recognizes that giving access to the training data or the underlying code and the internal workings of a system must be balanced with intellectual property rights, as the protection of trade secrets is a key staple of economic freedom.

The copyright issues surrounding the legal qualification of the code and its components, and its protection afforded by the copyright laws, are still being discussed today[10], although it is generally recognized.

For example, access to the code (or training data) does not guarantee an appropriate explanation of the inference logic used by the system. It could be incredibly complicated, used only in conjunction with other systems (thus only comprehensible by analysing both), or the training data could have been used for the development of a machine learning system, thus suffering from the black-box problem.

[10] In *Google LLC v. Oracle America, Inc*, the US Supreme Court held that Google's copying of 11,500 lines of declaring code--necessary to implement computer programs developed using the Java interface--was a permissible fair use because of the functional nature of the code being copied and the limited scope of Google's copying.

To summarize, the EU recognizes the inherent hardships involved in providing a useful explanation to algorithmic decision-making, and dismisses the problem by arguing that "Independent authorities or experts could be tasked with certifying and auditing processing methods or providing advice beforehand. Public authorities could grant certification, to be regularly reviewed".

The first issue that comes to mind regards the level of access that third-party authorities would need to assess and verify the level of compliance of a system.[11]

External auditing mechanism are usually conducted once a system has already been deployed, and given the extended complexity most systems have nowadays, and the fact that they are continuously updated, the auditing stage could last a long period of time, either making the process virtually useless or creating a bureaucratic or technical nightmare for the relevant experts.

This is due to the aforementioned issues related to the level of transparency that is to be expected of an AI system. While third party auditing companies exists, having external parties overlook the entire development process would give rise to legal problems related to copyright protection, solvable only through the use of non-disclosure agreements. Government supervision, or agencies controlled by authority institutions, seem like the only choice that would guarantee security as well as efficiency.

Although certain big name companies, such as Google or Microsoft, have published papers describing possible internal auditing standards, which can be compared to an airplane checklist, detailing steps to be followed throughout the development and deployment process. These mechanisms would provide benefits as they bypass internal security measures and the use of APIs to assess model outputs, but may lead to liability issues and lack of trust.

If a company is allowed to provide their own certification, the liability would then have to be assessed in Court on a case by case basis, and many smaller companies would find the risks too high to invest in research and development.

[11] Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. ID Raji et. al. 2020. FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Malta became the first government to commit to launching 6 AI public sector pilot projects in traffic management, education, tourism, healthcare, customer service and energy and water; based on an AI Certification Programme, to be under the management of an expert within the Malta Digital Innovation Authority[12].

Other countries such as Denmark[13], with its Data Ethics Seal, or the German Data Ethics Commission[14], have started to approach this problem from a similar angle, with clear directions and principles, but still a lack of practical guidelines.

As furthering AI research implies using massive quantities of diverse data, legislative bodies have tried to restrict and guide its collection, access, and use, in order to protect both the citizens and programmers of specific systems.

The first formal publication of a comprehensive rational standard guiding the principles of data use goes back to 2016. It theorized four fundamental concepts, appropriately named FAIR, which are findability, accessibility, interoperability, and reusability. Without going into detail on what each of these entails, it is important to note that the intent behind the conception of these principles is to apply them not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data, since all components of the research process must be available to ensure transparency, reproducibility, and reusability[15].

Overall, the FAIR guiding principles aim at providing unique identifiers for data and metadata, accurate indexing, easily accessible protocols for retrievability, a broadly applicable language for knowledge representation, and meeting domain-relevant community standards.

These aspects should act as guides for both human and machine-driven activities. Fairness in the use of data should be the founding step for ensuring accountability in the development of decision-mak-

[12] Malta Digital Innovation Authority, 2019.

[13] National Strategy for Artificial Intelligence, 2019.

[14] Federal Government's Data Ethics Commission ("Datenethikkommission"), 2019.

[15] FAWilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).IR Guiding Principles for scientific data management and stewardship', *Scientific Data*, 2016.

ing systems, and new tools and job areas for this purpose are being created by the EU regulations.

For example, the GDPR achieves two main purposes in the context of fair AI decision-making. First, it governs the processing of personal data, second, it establishes in article 22 a basic principle limiting the impact that fully automated decision-making systems have on the person's rights.

The article states that the data subject shall have the right not to be adversely impacted by a decision based solely on automated processing producing legal effects.

This norm is clearly to be read as an assurance for anyone not trusting (with good reason) decision-making systems in being able to provide a justification.

It is up to the human in control of such system to reason backwards, using the decision as a starting point from which to form his own conclusion. As it is up to the user to provide a reasonable answer, a simple ‘the system said so’ shall not be sufficient, but the human in control needs to understand the final answer, how that answer was reached and why the system has reasoned so, in order to be able to explain the conclusion in case of an appeal. This is the reason why fairness and explainability are fundamental principles in the development of AI systems, as when they are successfully fulfilled the user in control will have a much easier time constructing his own argument and justification.

3 Bias in AI

In recent sentences, Courts have dealt with many issues that stem from the discriminatory results of IT systems, especially regarding the correctness of the decision itself.

There are many kinds of biases occurring in the processing of data which can result in unfavourable results, such as historical, representative or aggregation bias. Misleading facts, poor or lacking representation, wrong assumptions, all the above may lead, often unpredictably, to unfair and biased decisions.

There are many kinds of discrimination when it comes to algorithms, we might classify types of discrimination into direct and indirect, systemic and statistical, explainable and unexplainable[16].

Direct discrimination happens when an individual's attribute or characteristic, usually protected under the law, is explicitly the cause of an unfair outcome.

Indirect discrimination is when individuals appear to be treated based on neutral attributes, however, as a result of a seemingly impartial element, it results in an unfair treatment.

When the discriminatory practice is inherent to the overall decision-making process of an authority or institution, that is systemic discrimination, while if the latter make their decision based on the average group statistics to judge an individual belonging to that group, it refers to statistical discrimination.

Explainable and unexplainable discrimination refer to whether, if a group or individual is treated differently from another, a reasoning for such discrepancy has been given. If the difference has been voluntarily introduced in order to counterbalance a previously discriminatory dataset it shall usually be deemed as a legal practice, while if it was not justified it is to be considered as illegal.

The most (in)famous case related to a discriminatory algorithm is the COMPAS case.

3.1 COMPAS

While the European Courts have not seen a widespread use of predictive algorithms, the same cannot be said for American ones. This paragraph shall examine the judgment of the Supreme Court in *State v Loomis*[17], where the plaintiff argued against the use of the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. This system uses machine learning to determine the risk for any subject to commit the same offence twice, and the system estimates the risk of recidivism based on both an interview with the offender and information from the offender’s criminal history. Being a privately owned system, the full training data and the algorithm used to train the ML system is protected under trade-secret, thus there

[16] A Survey on Bias and Fairness in Machine Learning, Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan, 2019

[17] State v. Loomis - 2016 WI 68, 371 Wis. 2d 235, 881 N.W.2d 749

is no real explanation as to why it gave a lower score instead of a higher one.

The United States Supreme Court had to establish whether it is a violation of a defendant's constitutional right to due process for a trial court to rely on the risk assessment results provided by COMPAS, preventing a defendant from challenging the accuracy and scientific validity of the risk assessment. Without entering into the full debate behind the final decision, it is important to note that the Court ruled in favour of allowing the use of COMPAS in the judiciary system due to the fact that the final decision was not based entirely on the score given by the system.

The machine learning component of COMPAS has been accused of having learned from biased data[18]. The reason for this stems from the fact that the US legal system is historically discriminatory towards people from minority backgrounds. This precedent conduct could have instilled a bias in the data, and therefore in the results of the AI. When those judgments are used to build a predictive system, the AI will unknowingly incur in those same human biases and therefore result in decisions heavily disadvantageous for certain groups of people. Even if the name of the person and its background is hidden from the system, the exclusion of a personal attribute is not enough, as the AI can recognize similar elements such as the area of residence, average income, job, and will infer personal and sensitive data from seemingly unbiased attributes.

An automated system is able to look through an large amount of information/data in an low amount of time, faster than any human ever could. We should then ask if it is possible for a human to review and assess the machine’s decision. Even if we had access to the full reasoning process of COMPAS, the systems looks at thousands of past cases in order to determine the risk of recidivism, something that a human would need months to do. Can we achieve both the respect of the right of a user to be accurately informed of the steps taken to reach that decision, as well as keep the biggest advantages of using artificial intelligence systems (quickness and efficiency)? While the

[18] There’s software used across the country to predict future criminals. And it’s biased against blacks *by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica* May 23, 2016.

two are not mutually exclusive, this paper shall later describe why explaining a black-box algorithm is not a task easily accomplished.

3.2 The Deliveroo Case

While COMPAS was based on a machine learning system, the following is, although not confirmed, thought to be a traditional symbolic algorithm.

Deliveroo is a London based online food delivery service that coordinates riders and participating restaurants.

The recent Deliveroo case saw the Italian court recognize that the algorithm used by the company was unfair. It worked by scoring a rider based on the attribution of a reputation score determined by the work performance of the previous week. The score was calculated considering two attributes: reliability and participation. The first index expressed the number of occasions in which the rider, after having booked a session, was actually present in a predetermined spot, ready to receive orders within 15 minutes from the start of the session. The second index instead considered the number of times the rider declared himself available for specific peak delivery hours.

After having registered for a session, the rider can cancel for a set amount of time.

According to the court, the system discriminated against riders who missed or cancelled the session (after the grace period) for legitimate reasons and were therefore undeserving of a penalty. Examples of these situations were handicap causing illness, family conditions, or the participation in trade union events. In all these cases, according to the applicants, the riders were exposed to a penalty in the score and a consequent relegation in the ranking, thus being penalized in the choice of shifts, causing in a loss of future job opportunities.

This case shows that, even when the algorithm's reasoning derives strictly from man-made rules, the data can introduce biases, on purpose or accidental. If the data is not translated correctly (and without biases), the computable law result may still end up being discriminatory.

This system suffered from indirect discrimination, as a neutral attribute such as the missed work hours, if not appropriately distinguished in the system, cannot consider the distinction between the different cases, while the law would account for reasonable abstentions and the motivation behind each lost hour.

This scoring system is no longer in use at Deliveroo, as it has been replaced with a more manual approach.

Proving that indirect discrimination exists can be a difficult task, especially in automated learning systems, for which even the developer could have difficulty in knowing exactly how it works. How could a judge or legal professional operate on the basis of such hidden information? In symbolic programming, if done correctly, it can be easier to spot such biases. The system's developer knows which attributes are taken into consideration by the system, and therefore can consider their correctness and validity.

Only once an explanation is given, can a system be judged and determined as biased or not.

3.3 Looking inside the Black-Box

In this section we will explore how can a machine learning system can provide an explanation. As previously described, it can be harder for a machine learning system to explain its reasoning. Various techniques and support machines have been developed, which aim at identifying the relevant features that a ML system takes into consideration when reaching an explanation, but rarely can those be useful in the legal domain.

For example, we find LIME (Local Interpretable Model-Agnostic Explanations[19]), FEATURE importance [20]and SHAP (SHapley Additive exPlanations). The principle behind most so called white-boxes (ML systems that mimic the original black-box system while providing an explanation), is based on adjusting the original input and analyzing the impact it has on the corresponding output. It then estimates the importance of the different input features in the output's generation, with respect to the given black box. To achieve this result in needs to work with an explainable representation of the input given to the system, as the final answer is a human-understandable list of explanations, reflecting how each feature weighs in on the final output.

[19] "Why Should I Trust You?" Explaining the Predictions of Any Classifier (KDD) by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 2016

[20] Comparison of feature importance measures as explanations for classification models. Mirka Saarela & Susanne Jauhiainen, 2021

It requires a vast amount of data to achieve an accurate result, and if the data used as input is not adequately structured, (like a series of numbers and weights, unpaired instances, etc.) the final justification will not be easily understandable. On the other hand, even if the data used is symbolic, the final motivation is simply an explanation of the internal reasoning process, that does not necessarily reflect an intelligent behaviour (for example, an algorithm measuring the length of the border of a picture, instead of the pixels making up the object itself, to classify it[21]).

This process is built upon the notion that a similar enough system using the same data set as the black-box one, will be able to approximately interpret the original by being trained on slightly perturbed inputs. While a precise enough copy, it is still only a copy and does not guarantee the complete accuracy of its results. The objective of the system and the reasoning used are entangled ideas in ML, and distinguishing one from the other is not something that can be done with high enough precision yet.

SHAP works based on the same principle but from a different perspective: rather than focusing on each individual feature, it balances the major ones that can be found in an output (such as the pixels of a picture) and compares the contribution of each one to the whole, through game theory. SHAP measures the relative importance of a feature by estimating its average contribution to the prediction, the Shapley value from coalitional game theory, by considering a feature as a player in a coalition of features.

Both models rely on faulty assumptions, as random adjusting of the input may take into consideration factors that the original machine did not, therefore create correlations which do not actually follow the original model; the human-made perturbations[22]may involuntarily lead to more bias. A potentially malicious user may also take advantage of such a system to forge an explanation by simulating perturbations in the data input and control the following explanation.

[21] Unmasking Clever Hans Predictors and Assessing What Machines Really Learn, 2019, Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller

[22] Perturbations refer to any changes in the input data.

The final method, “feature importance”, does not create an artificial white-box to explain the black-box, but it embeds an explainability mechanism inside the original system. It attributes a score to each feature of the input, and shows the relevance of each feature in the answer. Feature importance is similar to the first model, with the main difference being the focus of inherent the values used by the system, instead of the *ex-post*.

While each of these systems is useful in domains such as computer vision and scoring systems, it must be refined to be used in the legal system.

4 A Symbolic Approach

4.1 Knowledge and Rule-Based Systems

In the legal domain, the typical reasoning scheme is the application of rules[23]. Legal rules may be seen as conditional statements, linking an antecedent to a consequent so that from the former it is possible to infer the latter. Legal rules usually connect a set of abstract provisions of facts to a legal effect. To achieve this, the natural language needs to go through a process called normalisation. Normalisation is a method for representing legal information in such a way as to eliminate syntactic ambiguities. Its main aspect consists in substituting logical connectives to the connectors of natural language (like “and,” “or,” “if. . . then,” “unless,” and so on).

Through this process it is possible to describe how the legal norm can be written in a computable language, understandable by computers.

Systems based on the representation of knowledge, are called knowledge-based systems (hereinafter KBS) or expert systems (due to their reliance on human expertise). The process of extraction and formalization of the relevant rules is called knowledge engineering[24],

[23] See SARTOR, G., CASANOVAS, P., CASELLAS, N., RUBINO, R., *Computationals Models of the Law and ICT: State of the Art and Trends in European Research.* Springer, Berlin, Heidelberg (2008).

[24] According to Feigenbaum and McCorduck, knowledge engineering can be defined as following: “The art of bringing the principles and tools of AI research to bear on difficult applications problems requiring expert’s knowledge for their solutions. Knowledge engineering involves the cooperation of human experts in the do-

and the professionals responsible for the normalisation of these rules are called knowledge engineers. Any KBS has two essential components, the knowledge-base and a method for automatically applying and reasoning on that base, which is called inference engine, or reasoner[25]. The latter could be considered the 'mind' of the system, it is the instrument that interprets the rules and applies them to the facts given as input. In the legal context, the work of a reasoner could be compared to that of a judge, as the systems infers the epistemic consequences from the current factual situation, and reasons on the actions that must be implemented to solve any given problem based on that original situation.

In machine programming and deep learning systems the knowledge upon which the system reasons is implicit, mainly comprised of what the system has inferred in previous decisions. In order to manually add improvements to the reasoning of the system, such as adding a new case, human intervention is still needed, and if the change is big enough, a complete overhaul of the training data is needed. Often, traditional systems do not give the user any useful justification for the answer that has been reached. On the other hand, knowledge based systems, which are written in a language more understandable by humans (such as formal logic). The system involves a tailored inference engine, thus a way for the program to reason on the factual situation. The consequence is that the user has now available, together with the conclusion of the reasoning process, the premises that brought to that conclusion, as well as the specific inference rules that were. Any user that finds himself in the situation of having to update the system only needs to update the knowledge base, while the inference rules will remain the same.

Rule based systems (hereinafter RBS) are a type of KBS, as they involve a knowledge base and an inference engine. In particular, the knowledge base is made up of a set of rules. RBS are defined as being (1) transparent, as the rule base can provide a kind of explanation of its line of reasoning and answers to queries about its knowledge; (2) heuristic, as they reason with added knowledge as well as with formal

main to work with the knowledge engineer to codify and make explicit the rules (or other procedures) that a human expert uses to solve real problem".

[25] It is also called 'rule interpreter', by TURBAN et al., *Decision Support Systems and Intelligent Systems.* 7 a ed., Pearson Prentice Hall, (2005).

knowledge of established theories; and (3) flexible, as they can be enhanced to integrate new knowledge incrementally into its existing store of knowledge[26].

The most interesting area of study, when discussing the development of new RBS, is the representation of the knowledge base. Ever since the advent of expert systems, the basic requirements of any representation scheme have been the following: extendability, simplicity and explicitness[27]. The rulebase needs to be constructed in a way that allows extensions and upgrades, without being forced to revise the whole knowledge-base. In any given domain, and especially the legal one, the experience that comes with the progression of society gives way to new heuristics and forces distinctions between new ones and old ones. Laws are being rewritten constantly, and a system that would not allow for changes to the knowledge base is very limited in scope. Moreover, the most effective way for building a knowledge base is by implementing incremental improvement, as it can be updated and maintained in the future. Experts cannot define a complete knowledge base all at once for interesting problem areas, but they can define a subset and then refine it over many months or years of examining its consequences. All this argues for treating the knowledge base of an expert system as an open-ended set of facts and relations, keeping the knowledge itself as modular as possible.

The data structures that make up the knowledge base require conceptual simplicity and uniformity. While the usefulness this is easy to understand, it is nevertheless fundamental for the explanation of the contents of the knowledge-base, and of the analysis of links among the structures. Finally, the knowledge base needs to be understandable by programmers and users alike. The purpose of representing much of an expert's knowledge is to give the system a rich enough knowledge-base for high-performance problem solving, but because a knowledge-base must be built incrementally, it is necessary to provide means for inspecting and debugging it easily. With an explicit knowledge-base, it is easier for the experts who are building the system to determine what essential rules are present and (by inference) which are absent.

[26] See B. G. BUCHANAN e R. O. DUDA, *Principles of rule-based expert systems, in Advances in computers*, vol.XXII (1983).

[27] Ibid.

The development of legal knowledge-based systems is a complex issue. The reason lies in the fact that law is a complex social and normative system, it consists of both formal and informal logic, it has both theoretical and pragmatic components. Moreover, the complexity in the legal domain is determined by a large number of interacting components.

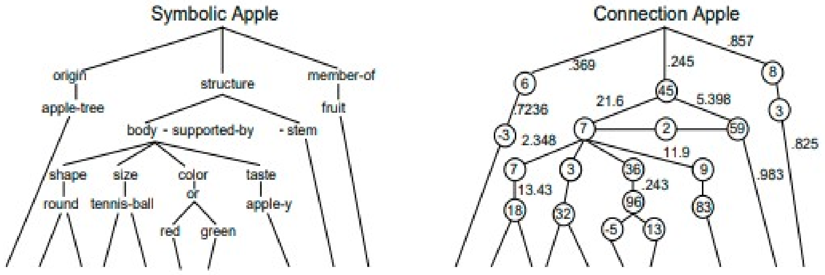
4.2 Opacity in Symbolic Logic Programming

In Symbolic AI systems, the problem of providing an explanation is expressed through human readable logic representations.

A key advantage of Symbolic AI is that the reasoning process can be easily explained, since it derives from a human representation. A Symbolic AI program can be made to explain (log) why a certain conclusion is reached and what the reasoning steps had been. This has the advantage of being human-accessible, but can show the limit of this model, since the knowledge has to be mostly hand-tailored by the engineer.

Machine learning systems, on the other hand, can extract from the source data that the engineer did not identify.

In the following picture, on the left, we can see how a human engineer needs to describe to an AI system an apple, in contrast with neural networks, on the right, that rely on a series of numbers and weights. This ‘symbolic apple’ was first introduced by Minsky in 1991[28], and since then a solution to the problem of opacity has been a constant focus for researchers.



[28] Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy, Marvin Minsky, In *Artificial Intelligence at MIT, Expanding Frontiers*, Patrick H. Winston (Ed.), Vol.1, MIT Press, 1990. Reprinted in AI Magazine, Summer 1991.

A similar reasoning is applied the computable representation of a legal source.

4.2.1 Requirements of a Knowledge-base

To convert legal norms in computable inference rules, and to provide the logical connections, it is important to achieve isomorphism, a one to one correspondence between norms in the formal model and the legal source text. There are several conditions that must be fulfilled, and they are following[29]: (i) Each legal source is presented separately; (ii) The representation preserves the structure of each legal source; (iii) The representation preserves the traditional mutual relation, references and connections between the legal sources; (iv) The representation of the legal sources and their mutual relations (...) is separate from all other parts of the model, notably representation of queries and facts management.

To build a knowledge-base the first step is to define an ontology, a representation of the concepts and their relations. The right tools need to be selected to enable the representation of the logic structure derived from the original legal text. Different programming languages and models may be adequate for different uses.

Programming languages commonly used in software development, such as Python or Java, are so-called imperative, and define a sequence of instructions that must be followed to reach a certain goal. On the other side of the spectrum there are declarative programming languages, such as Prolog, which express logic statements, that the interpreter should try and verify. It is possible to reach a goal by building a set of axioms and rules that closely follow deductive reasoning. The simplest way to explain the approach Prolog uses to reach a goal is to describe its "top-down" strategy: by starting from a high level, and attempting to verify the necessary premises and rules through propositional logic. It is exactly what law is supposed to be, a generalised statement from which singular behavioural patterns can be deduced. This basically reductionist technique is typical of the approach to AI called heuristic programming.

ML algorithms work in the opposite way, using a bottom-up approach. Starting with simpler elements - elementary logical principles

[29] Bench-Capon, T.; Coenen, F.: Isomorphism and legal knowledge based systems. Artificial Intelligence and Law 1/1, 65{86, 1992.

or simplified models and examples - and then moving upwards in complexity by finding ways to interconnect those units to produce larger scale phenomena.

4.2.2 Prolog and CrossJustice

Symbolic approaches have been implemented for the computable representation of sources of law, in many different domains.

For example, the CrossJustice platform, which has been developed by researchers from all over Europe, including the University of Bologna, makes use of Prolog to represent the European directives and national implementations in the field of criminal justice to create a decision support system. The Project aims to identify the critical gaps in the implementation of the European directives, and to provide solutions in a comparative perspective, in order to improve the efficiency of judicial systems and their cooperation, thanks to information and communication technology. Furthermore, it focuses on the compliance of national instruments implementing EU directives with the EU acquis, as well as the compatibility between national frameworks as resulting from the implementation of EU directives. To this end, the CrossJustice online platform aims at providing an advisory service on the effectiveness of procedural rights mainly directed to legal professionals, but accessible to law students, NGOs and all EU citizens. The system makes use of the Prolog language, which uses propositional logic, and assumes universally quantified variables to create rules to prove that at least one instance exists.

5 Representing fairness

A logic-based approach inherently provides trust in the system.

A computable, symbolic, representation of a source of law provides the legal practitioner with an inference process based on the logic relationships between the facts of the case and the relevant articles that shall be applied. This is similar to the reasoning the legal expert would make. The user can see why a result was reached, how the machines has reasoned, the facts used by the system, the relevance of each fact to the final answer, and the system can be trusted to have reached a fair and non-discriminatory conclusion as it is based on the original source of law.

It is immediately explicit to the user whether the system has based its answer on race, gender, or any other discriminatory feature, as that fact would have been visible in the reasoning process.

Furthermore, symbolic approaches provide an easier *ex-post* inquiry on the systemic or indirect discrimination of a seemingly fair system, through the use of a meta-interpreter. It is an added program, structured in the rulebase code, that generates a proof tree tracing the execution of the program, by printing a log of the evaluated predicates. The system provides an user-friendly explanation of the facts of the case, and the trace/debug modes further reinforce the explainability of the program by logging the transition through the resolution states of the predicates. A Prolog trace is a sequence of events which gives a picture of a program execution. To summarize, the program is able to document and log the facts the system recognizes as true and are needed to reach the solution, the rules that are applied and the logical inference process the system makes through the different rules.

Assuming a set of rules exists, Prolog implements logical formulas and constants to represent elements, sets (which are made up of elements) and binary relations between sets. The system is able to look through the facts and, from those, infer all new possible elements. It does not lead the user from one fact to another, but it takes into consideration all the given information and, through deductive reasoning, reach the solution. While the result is essentially the same, the reasoning behind the procedure offers a more realistic interpretability.

The user is able to see why a rule was not applied, which elements were taken into consideration before others, the elements ignored by the system and the overall inference process, from the facts to the final result. By providing slightly different inputs and following the reasoning of the system, it becomes much easier to see whether the system changes its answer, and why. This is the reason why the algorithm used for scoring Deliveroo riders was quickly found to be discriminatory.

On the other hand, symbolic programming can be a complex and long task, especially in a vast domain. One of the reasons why machine earlier has been increasingly popular is the possibility of expanding the system by learning from data, instead of direct input from the knowledge engineer. A symbolic knowledge base needs to be manually kept up-to-date and accurate, and the human programmers need to add all relevant sources for the system to be complete, which in the ever-changing legal domain does not appear to be feasible.

Neuro-Symbolic integration research aims at combining the advantages of both machine learning and symbolic AI systems. Using symbolic knowledge to add constraints to the machine learning algorithm, adding the predictive and classification ability of the latter to the explainability of the former. Techniques such as KBANN[30] (Knowledge-based artificial neural network), DNN[31] (Deep neural network) with Logic Rules and LTN[32] (Logic Tensor networks), are all learning methods that model a symbolic knowledge base in order to teach a machine learning system the data and logic constraints that have been embedded, thus enabling the system to predict new facts consistently according to the logical norms.

Such works are discriminated inside the model integration category depending on the kind of logics they leverage upon: "logic and numerical" integration vs "numerical, statistical, and logic". On the other side, the model composition category is split depending on the kind of composition: symbolic knowledge extraction vs symbolic knowledge injection. The former subcategory includes those approaches where some sort of symbolic knowledge is somehow extracted from sub-symbolic models - namely, rules, and tree extractors - whereas the latter includes those approaches where some sort of symbolic knowledge is injected into sub-symbolic models[33].

[30] G.G. Towell, J.W.Shavlik and M.O.Noordeweir, Refinement of Approximate Domain Theories by Knowledge-Based Neural Networks, in: *Proceedings of the Eighth National Conference on Artificial Intelligence*, 1990,pp.861–866. https://www.aaai.org/Library/AAAI/1990/ aaai90-129.php.

[31] Z. Hu, etal., Harnessing Deep Neural Networks with Logic Rules, in: *Proceedingsofthe54th Annual Meeting of the Association for Computational Linguistics (Volume1:LongPapers)*, 2016,pp.2410–2420.

[32] M. Abadi, etal., Tensorflow: A system for large scale machine learning, in: *12th* USENIX *Symposium on Operating Systems Design and Implementation*, 2016,pp.265–283.

[33] Calegari, Roberta, Ciatto, Giovanni, and Omicini, Andrea. 'On the Integration of Symbolic and Sub-symbolic Techniques for XAI: A Survey'. 1 Jan. 2020 : 7 – 32.

Creating a symbolic knowledge-base, encompassing the representation of the various sources of law, be they legal or regulatory acts, case-law and legal argumentation theories, will exponentially increase the capacity of these systems to provide an explanation to their own reasoning processes.

Although the field has not seen much impact in the legal domain, it seems like the way to go forward.