



Bias and Inference in Neural Networks and their relationship with the Law

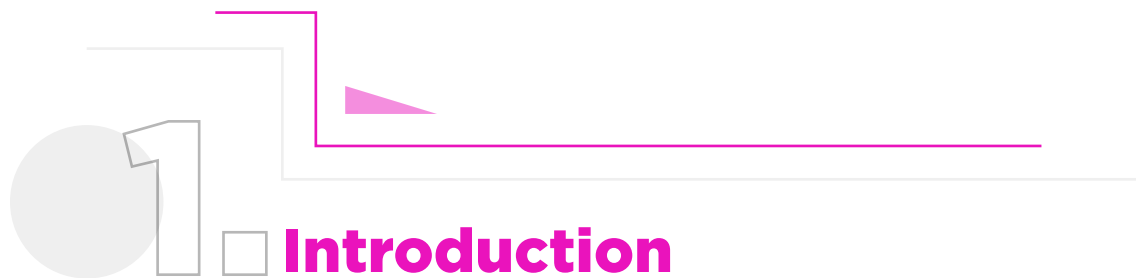
Carlos Amunátegui Perelló ^{1*}
Raúl Madrid ^{2**}

Abstract

The objective of this article is to approach the bias phenomenon generated through neural networks, whether in its training or in the design of their objective function, and to analyze some potential implications in the legal domain.

^{1*} Professor of Artificial Intelligence, Pontificia Universidad Católica de Chile. Researcher of the Science and Technology Program of the Law School.

^{2**} Director of the Law, Science and Technology Program, Pontifical Catholic University of Chile Law School.



1 Introduction

Based on the first deep neural networks, developed in 1987 (Rumelhart et al., 1986, 533-536.), the development of artificial intelligence systems based on such networks has become increasingly dependent on the mass accumulation of data which feed the algorithms designed (Huang et al., 2013), to such an extent that now a high volume of data feed is more relevant than its elegance or efficiency (Lee, 2018, 14).

Due to this growing dependence, the so-called “bias” problem has worsened. The Spanish word “sesgo” has been used to translate the concept of bias, although, in fact, in Spanish the term for “prejudice” would fit better. The word “bias” is defined by the Cambridge English Dictionary as “a situation in which you support or oppose someone or something in an unfair way because you are influenced by your personal opinions.”³ The pejorative meaning is evident if compared with the definition given by Real Academia de la Lengua Española, which offers the word “prejuizar” (prejudge): “judge someone or something before appropriate or without having full knowledge about them.”⁴ In both cases this is a loathsome action either because it is done with willful misconduct, with an (apparent) mental impairment or simply with a lack of liability. The term “sesgo” (bias) is more subtle to express the same idea in general because as an adjective it means something twisted, cut or placed obliquely (Barcia, 1889). This idea related to obliqueness is also present in the English “bias”, as it refers to an error when judging something, but preceding the word “prejudice” (Roget and Davidson, 2003, 204).

³ Cambridge English Dictionary, voice “bias”. Available in: <https://dictionary.cambridge.org/-dictionary/english/bias> The translation is ours.

⁴ Real Academia de la Lengua Española Dictionary, voice “prejuizar”.

From our research perspective, the concept of “sesgo” (bias) points exactly at the possibility of there being a prejudice in the meaning mentioned above, every time the data accumulated feeding algorithms may be affected by all kind of presumptions, preventions or preconceptions, either conscious or unconscious. These preconceived ideas may derive from information gathering or accumulation process or from algorithm design, specifically from its objective function, which may result in cases of arbitrary discrimination made by the algorithm, even if those who have designed it are not aware of that bias.

In fact, from a legal perspective, the problem is deeper every time that the Law not only limits the ability of legal operators to make decisions based on certain criteria (like race⁵, for instance), but also it sometimes leads to making decisions based on explicitly formulated criteria, such as dangerousness or the possibility of escape during the temporary release⁶.

Additionally, there are problems and incompatibilities that are generated between neural networks and the legal system in general as a body of rules. In this sense, we will try to elucidate the particularities of bias in neural networks and its possible causes, and then establish the problems and incompatibilities that are generated between these and the legal system in general as a body of rules.

⁵ Thus, the Constitution of the Republic (CPR in Spanish), in article 19 No. 2 sets forth the right to non-arbitrary discrimination. This right has been developed in many aspects by Law 20.609, which second article establishes the following definition for arbitrary discrimination: "Article 2nd - Definition of arbitrary discrimination. For the purpose of this law, arbitrary discrimination shall mean all distinction, exclusion or restriction that has no reasonable justification, made by State agents or individuals and which causes loss, disruption or threat to the lawful exercise of the essential rights provided for by the Constitution of the Republic or the international treaties on human rights ratified by Chile and currently enforced, especially when they are based on reasons such as race or ethnicity, nationality, socioeconomic situation, language, ideology or political opinion, religion or belief, union activity or participation or lack thereof, gender, motherhood, maternal breastfeeding, sexual orientation, gender identity, marital status, age, filiation, personal appearance, and disease or disability."

This criterion has been confirmed by several rulings by the Supreme Court (CS in Spanish), among which we can mention the recent one issued on October 23rd 2017 in the case 'Huerta vs Sociedad Plaza', Rol 2847-2017, sixth clause:

'[L]ife within society implies that there will be differentiation because there are choices determined on a daily basis; however, the law sets forth arbitrary discrimination, i.e., distinction lacking in rationality, which is just justified by the whimsical attitude of the one it is performed by.'

It is in this context that section 2nd hereinbefore delivers guiding criteria by which to set an arbitrary discrimination –disability, among others. In effect, any distinction, exclusion or restriction of rights made between individuals on the basis of disability, which is not supported on reason, is the action punished by the law."

⁶ (CPR 19, No.7 e) 'The accused shall be released unless arrest or preventive custody is considered by the judge as necessary for the investigations or for keeping the injured party or society safe. The law shall set forth the requirements and manners for this.'

2

Neural Networks and Correlations

Generally speaking, an artificial neural network is a network of simple elements (nodes) organized into an interconnected hierarchy, massively and in parallel, which seeks to interact with the objects in the real world just as does the biological nervous system (Gurney, 1997, 13). In principle, neural networks are configured by programmers, who set up the model architecture. Then, these networks are trained in a set of data, through which the model determines the weights it will assign to the different connections between layers, and to a certain extent, it configures its hidden layers. Thus, to function efficiently and detect correlations between the data it is trained with, a neural network needs a great number of examples. Thus, the algorithms formulated on the basis of neural networks depend largely on the data feeding them. Simply put, a neural network is trained with a number of examples taken from a database, from which it will correct the weights of its axons through backpropagation, usually following an inverse stochastic gradient descent of the integral equation⁷. Whatever the method adopted, which can be the classic supervised learning or an unsupervised one, or even generative adversarial networks (GANs)⁸, neural networks will always depend on the quality of the data they are fed with, since it is based on these data that the models generated will become efficient and effective.

⁷ See: Hinton and Salakhutdinov, 2006, 504-507.

⁸ See: Radford et al., 2016.

⁹ We refer to the seven largest companies that dominate the web and have the largest investments in artificial intelligence, namely Alphabet (Google's parent company), Alibaba, Amazon, Apple, Facebook, Baidu and Tencent, which have a dominant position in data accumulation and web traffic.

If neural networks did not work well until 2010, it was mainly due to the relatively poor data available. This was compensated by the huge information bases generated through the Internet (Kai-Fu, 2018). Thus, the current artificial intelligence technologies are fully dependent on data and their efficacy is significantly determined by them. In fact, the large availability of data generated by the Internet has been key in making the current neural network models effective, since thanks to the abundance of thousands of photos, videos, writings and searches the current algorithms are trained and the weights are configured, bonding their different layers. Nothing is as efficient as having more and more data, so that even a poorly designed algorithm trained in a huge database works better, i.e., has more predictive capability than a more refined one that has been trained with a smaller database. Having hundreds of thousands of million data of different kinds has allowed Internet⁹ companies to develop and put into practice this technology.

The expansion of the Internet of things has also boosted available data, so nowadays data generation covers much more than before, ranging from the images taken with car cameras, geolocation by GPS to all the data issued by devices connected to the Internet in general, whatever their kind. Since the abundance of data optimizes algorithms, the larger the amount of data available for companies, the better algorithms they will have, and therefore, the better products with artificial intelligence functions they will have as well¹⁰. The concentration of data has given rise to tensions since it paves the way for a natural oligopoly in technology between the largest Internet companies.

It is evident that this dependence on data means a greater problem when it comes to assessing the results presented by such models, because the quality of their results seems to depend on the data with which those networks are trained. Thus, if a network is trained with data that show some kind of bias (racial, sexual or other), the results generated by the network will show the same bias, not only reproducing it, but also even amplifying it¹¹ and institutionalizing it in the sphere where it is applied.



¹⁰ This is what has been called the Matthew effect, since those who have more will receive more. See: Pasquale (2015, 82).

¹¹ See: O'Neil (2016, 23).

The biases that a model may take from the data with which it is trained are usually classified into at least three types: interaction bias, latent bias, and selection bias. In the first case, without noticing the user or programmer introduces a bias in the model through the way in which they interact with it, for example, when the objective function of the model or the object searched is defined.

The latent bias takes place when the model makes inappropriate correlations, generally when setting false links between data points. Thus, for instance, if people who are hard-up do not pay their credits, and the lack of creditworthiness can be correlated with poverty, which can be correlated with spatial segregation in the city, an algorithm can take someone's residence in a segregated city point as an indicator of high credit risk –likely to be completely false. A selection bias takes place when the database is not representative enough of the diversity in a certain social environment. For example, if an algorithm is trained to determine elements with which to predict soccer-playing abilities in a given population good using medical data of all the first-division players in the Argentine Soccer League, that algorithm is likely to be useless to make this prediction in Japan, due to its low representativeness of Asian population.

When it comes to legally assessing these applications, the bias problem we have just described turns out to be very relevant. As has already been pointed out, an algorithm has the quality given to it by the data with which it is trained; the only thing that neural networks do is to establish correlations between data, but the conclusions reached will depend on the quality of the data provided. For example, if an algorithm must be trained to detect different dog breeds, and domestic dogs always appear in household contexts while wolves appear in natural environments, the mere fact of having a dog in a wild environment (a snow covered forest) will lead the neural network to associate it with wolves, as has happened before¹². There are many cases in which an algorithm makes mistakes that are implicit in the data from which it is trained to make the correlations.

Some cases are a real cause of concern. An example for this is what happened with the facial recognition program for Facebook. Joy Buolamwini (Buolamwini and Gebru, 2018) determined that the algorithm for face recognition used by this company was deficient to recognize black people's faces –it simply did not detect them. Going further back to 2015, Google's facial recognition program had identified two black men as gorillas (Zhang, 2015), for which the company had to publicly apologize. These cases seem to have been due to the fact that the database from which Google and Facebook train their algorithms did not have enough data on black people, causing a serious problem when applied.

Nevertheless, there are more serious problems related to potential biases an algorithm may contain. An example for this can be found in the case of the COMPAS recidivism prediction algorithm used in some places in the United States of America. Since arbitrary discrimination is a major problem all over the world, some states in that country started to use an algorithm that gives judges recommendations on their sentencing and on temporary releases. This takes into account the risk of recidivism of those accused or convicted in each case. The objective was to generate decisions based on data and, therefore, detached from the risks derived from human prejudices. However, the result was frankly disquieting. Taking into account psychological questionnaires filled out by inmates, the algorithm systematically recommended longer penalties and ruled against granting temporary releases to colored people, even if they had committed the same crime and had the same criminal record as whites.

¹² The issue occurred in 2017 and shows the fragility of the correlations made by the neural networks. See a press article analyzing the case: <http://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learning-model-to-avoid-adoption-errors/>

In other words, when the judge had to decide whether to grant a temporary release, it made a distinction by race, a universally prohibited criterion¹³ which is considered discriminatory and arbitrary¹⁴. It is very interesting to see that the result obtained is exactly the opposite to the result sought, since instead of reaching blind justice, the blindfold covering the Lady's eyes seemed to have fallen, affecting her impartiality. As we have said before, the problem seems to lie in the fact that algorithms get the information and establish correlations based on the data that they are fed with, so if these data are biased, which may often derive from historical conditionings, that bias will be reproduced by the model. It may even be amplified, with implications in the future¹⁵.

¹³ Thus, the Universal Declaration of Human Rights of 1948 sets forth in Article 2:

"Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty.

¹⁴ The investigation on the case was uncovered by Propublica, a non-profit organization which, among other things, analyzes the use of artificial intelligence. Available at: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>. Visited on June 12th 2019

Another interesting case is the selection of CVs. Since it is a hard task, especially in areas with lots of applicants, it is possible to automatize this job to shortlist candidates, one of whom will be selected. A commonly used criterion to design these algorithms is to consider as good employees those who have been in the company for some years and have been promoted. The CVs received are compared with those models. The results have failed on many occasions, since in several companies, especially in IT, men prevail over women, so algorithms tend to discard CVs from women who apply for those positions just for being women¹⁶.

¹⁵ El asunto resultó en una demanda y fue determinada la existencia de sesgo en el algoritmo. Véase: *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016). El caso se encuentra analizado en: *Recent Cases*, en *Harvard Law Review*, 130, 2017, pp. 1530-1537.

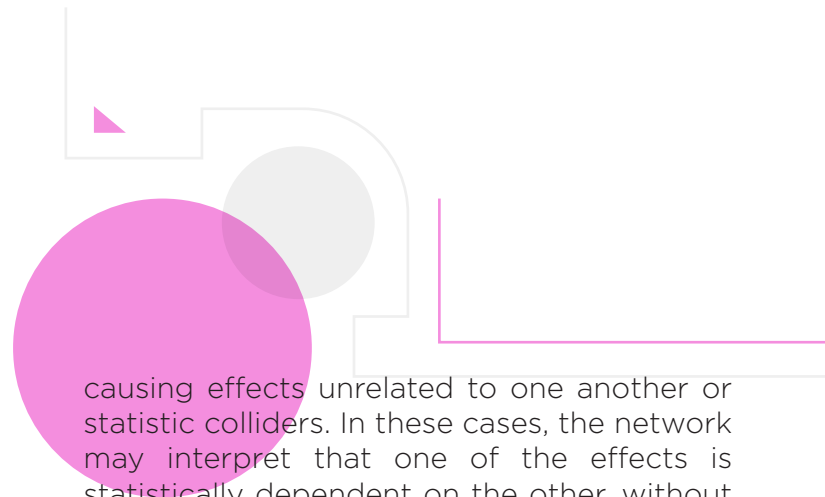
¹⁶ Los casos en que esto ha ocurrido son muchos. Como muestra, véase el reportaje de Reuters respecto a Amazon, disponible en: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> Consultado el 12 de junio de 2019.

¹⁷ El ejercicio fue diseñado y analizado por Meredith Broussard. Vid: *Broussard* (2018, l. 2150).



from the Titanic¹⁷. It is a simple formulation exercise of a neural algorithm that is often recommended for programming beginners. Based on a list of passengers who survived in the Titanic accident and data taken from the original records, a model is designed to predict the probabilities to survive a catastrophe at sea. The result is that the most relevant factors to survive the collision with the iceberg are being a woman and travelling first class. Therefore, if we design a predictive algorithm based on this model for an insurance company for instance, poor men would pay much more because they pose a higher risk in maritime transport than do well-off women. This kind of algorithms are relatively simple to make and are being used in areas like health insurance and bank credits despite our unawareness, which is worrying due to the underlying arbitrariness¹⁸. Should poor people pay higher interests just because they are poor, even if they have a spotless credit score?

A major problem with neural network systems is that their results are always long correlation games. Even if the databases used are suitable and do not carry significant biases, whenever they establish a correlation between data points and results, the correlation established is likely to contain some confounding. This might be due to multiple reasons, such as the same factor



causing effects unrelated to one another or statistic colliders. In these cases, the network may interpret that one of the effects is statistically dependent on the other, without taking into account that both depend on a third one¹⁹. Thus, in a highly segregating environment, some proper nouns in certain minorities may be related to poverty, while extreme poverty could be related to the lack of purchasing power, and the latter to a higher insolvency risk. Not having a causal reasoning and not being able to distinguish social processes suitably, the neural network could correlate certain proper nouns with insolvency, and deny access to credit to people bearing those names, regardless of their purchasing power, credit score or real insolvency risk. In short, the presence of a bias in neural network systems is not just a problem related to data quality but possibly also to a structural risk from an architecture that has not been designed to detect causes and effects.

¹⁷ The exercise was designed and analyzed by Meredith Broussard. See: Broussard (2018, I. 2150).

¹⁸ Frank Pasquale (2015, 38) gives us some grim prospects: "Reputation systems are creating new (and largely invisible) minorities, disfavored due to error or unfairness. Algorithms are not immune from the fundamental problem of discrimination in which negative and baseless assumptions congeal into prejudice"

¹⁹ On this subject, see: Pearl and Mackenzie (2018, 135 and following).

3

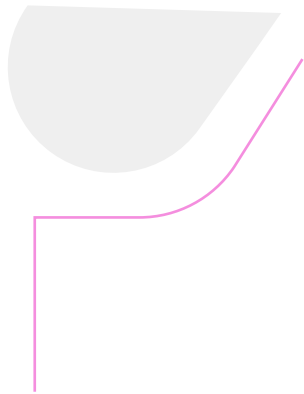
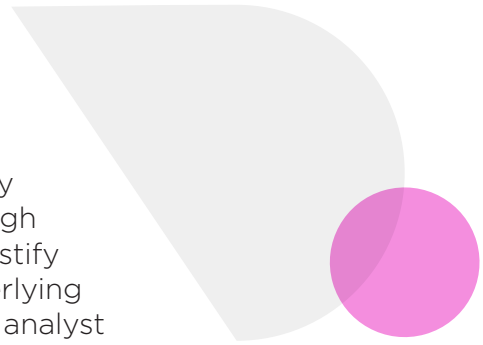
Chilean Law and Correlations

This presents us with a serious problem: the Law is a regulated inference system. It can assume that a seller must comply with the obligation to deliver the item sold on the basis of some rules preset in the Civil Code, like the existence of a purchase agreement, the degree of compliance by the buyer to meet their own obligation to pay the price, any vices, among others. An algorithm may well infer the existence of the obligation to deliver the item (or the lack thereof), based on any other factor included in the data with which it was trained, such as for instance the name of one of the parties, which the algorithm may find extremely important given the data it has. Strange as though it may be, it is perfectly possible²⁰. If this algorithm has been designed to recommend a judge to pass a sentence or to decide how a conflict should be settled, the result may well be illegal and arbitrary. In principle, this could be prevented with better data or directly by discarding the recommendations made by algorithms that are not based on lawful correlations. Yet this is not easy to determine since the neural network models tend to be opaque and not highly self-explanatory.

²⁰ To make this example more explicit, we can imagine a company that on several occasions has been sued for having failed to comply with its obligation to deliver the item sold, for example a multiple store, of which the algorithm correlates the name of the multiple store sued with the sentence to deliver the item sold, although in the particular case the legally necessary assumptions for this purpose do not exist.



In principle, neural networks are not able to explain their own predictions themselves. Up to now they are not able to compellingly explain how their decision is grounded. Unlike the old expert systems based on logical principles²¹, there is no perfectly traceable and understandable assumption chain, although the expert legal systems did not have the tools to justify their conclusions, not just e. To identify the reasons underlying a certain determination using neural networks, a data analyst should establish which neural connection is being considered relevant. Given their complexity, in some cases even getting to thousands of hidden layers, this might be impossible. In other words, the factor that can justify the action is still a human being, equipped with all the cognitive and interpretative components typical of a specific individual.



This opens the door to a new issue related to the previous one: the opacity of the deep learning systems based on neural networks. So far there is no simple mechanism to precisely determine the correlations made by an algorithm and the strength that each element that may come up in a layer from a network will have in the end. To do this the job of data analysts is required and, even in this case, it may be hard or practically impossible. Very often in order to determine whether a model has a major bias, the simplest thing to do is to apply it to a case in which said bias does not exist and to another one in which it does, and then to see if both cases are treated the same way –although there is a high degree of uncertainty in this procedure. It gets even more complex than this if we consider that the companies that design or use algorithms do not usually share them since they are the trade secret that allows them to get a significant commercial traffic and. Should the algorithms be made public, companies would lose value very quickly. In a nutshell, these algorithms are not only opaque when it comes to decision taking, but the way they work is often a trade secret.

²¹ We refer to the connectivist systems that prevailed from the 1970s until the development of modern neural networks. Such models reduce decision problems to a set of inferences made from manually programmed heuristic rules. These systems are still in use, although they had their hegemony in the 1980s when the so-called "expert systems" were developed. Also known as GOFAI (good old fashion AI).

In its Ethics Guidelines for Trustworthy AI, published on April 9th 2019, the European Union requests the artificial intelligence system design to be guided by the principle of explicability, so that algorithms should be as transparent as possible when it comes to the decisions they take. However, it is not clear how technically possible this is. This problem is currently being worked on, and there are several alternatives to it, although how far this path can lead us is not certain at this moment.

Some authors estimate that artificial intelligence will become a major problem for human rights in this century (Noble, 2018, 1). This becomes even more relevant when the field in which artificial intelligence is to be applied is not merely contracts but while performing a public function, like the administration of justice or the provision of social services. In principle, the actions taken in the public sector must be based on preset, objective rules, treating all individuals alike. This is what equality before the law

consists of. It is a duty set forth by the Constitution for all actions by the State (19 No. 2 CPR). This equality also implies a non-discriminatory treatment not only by public bodies, but also by individuals when relating to one another. This way, any differences that cannot be reasonably justified are against the political-legal order. Can an essentially unexplainable mechanism meet these duties implied in the idea of equality before the law and prohibition of discrimination? What's more, can an algorithm (whose decisions are based on data correlations that are dark for the individual receiving the law and for the individual imparting the law) take decisions that meet the basic legal grounds and transparency requirements demanded by our legal order to act in public bodies?

An algorithm is not smarter than the data with which it is fed²² and if these are biased, then the algorithm will get the bias from the data with which it has been trained. This is a problem related to overgeneralization or overfitting –the statistical terms to define it. The algorithm detects patterns in data, which on occasions are invisible even to the programmer, and then reproduces and amplifies them (Surden, 2014, 106).

A second relevant element is that the decisions of the algorithm are determined by its definition of success. And since this has been determined a priori by model designers, if the biases from the past have not been corrected, the model will simply reproduce them. To correct the discrimination problems that a set of data may show, it is necessary for programmers not only to be aware of their existence but also to actively correct them. Otherwise, the so-called objectivity they pretend to achieve will not be attained and, what's more, it is highly likely for the model to stress the biases. The agent will just reproduce the biases found on the data, producing results

²² In other words, humans are smarter than data. See: Pearl and Mackenzie (2018, 21).



that are morally questionable and legally unacceptable. A disturbing element is related to the price policy and employment conditions that such algorithms may generate. Setting profiles for different users, it is possible to establish different prices for the different individuals seeking a service. Already in 2000 Amazon admitted to using this policy (Broussard, 2018, l. 2128) –and this seems to be a fairly common practice in the cyberspace.

An effect caused by this mechanism is that people with a higher purchasing power tend to get lower price deals than those with a lower income (Broussard, 2018, l. 2128). Based on the searches conducted by an individual, on the opinions they leave on social network or any other set of data, an algorithm can be designed to detect a specific individual's potential needs and send them personalized offers under different conditions from those offered to other netizens. Google Mail's AdSense service detects patterns in the content of emails that each user receives in order to show them personalized ads. Apart

from the problems that such a service may entail since it violates email privacy (Chopra and White, 2011, 110), this mechanism generates a user profile and offers goods and services targeted to them specifically from a set of data.

The question at this point is if those offers may be affected by any bias, and if the differentiated price policy complies with consumer protection²³. As to the first problem, the potential existence of a bias in the generation of a price policy will depend on the case and we cannot give an a priori response. Since algorithm decisions are like black boxes, we cannot be sure of the reasons for a model in particular to segregate its price policy. The database may contain a bias or the objective function may be incorrectly designed, in which case the result will be discriminatory. The important thing is to analyze the case that is considered suspicious and to compare the results from the algorithm if the element that might be causing discrimination is incorporated or excluded. It is also important to pay attention to the major new trends when studying the prices offered. Evidently, if the results can be explained by arbitrary differences like gender, race or any other similar factor, that will be a case of algorithm-driven discrimination and the possibility to exercise constitutional acts like the protection remedy in Chile or consumer protection²⁴. However, due to the dark nature²⁵ of these algorithms, a specialized analysis should be conducted. It must be remembered that an algorithm may be discriminatory even if it does not take the factor that gives rise to the discrimination among its data points. In the case of the model that selected CVs in Amazon²⁵, gender was not among the elements considered by the algorithm. Its application resulted in the exclusion of women every time the algorithm positively considered the activities often performed by men, such as being part of a football team, and every time it considered as neutral or negative those activities often performed by women, like being a cheerleader for instance. Hence, although the agent did not explicitly consider gender, the result was gender-discriminatory.

Now supposing that the algorithm price policy is based on priori non-discriminatory criteria, the question that remains open is whether it is acceptable for different consumers to receive different prices. In theory, there is a (higher) general price which would be offered to most consumers, but in practice, given that a lot of consumers receive different prices when searching products, the general price does not seem to exist or if it does, it just for a minority. This seems to go against the provisions

²³ This problem was addressed by Zuiderveen Borgesius (2018) in a report to the Council of Europe.

²⁴ It would violate Article 3(c) of Law 19496 which sets out the right of the consumer not to be arbitrarily discriminated against.

²⁵ This example is also recovered in the article by Gómez Mont, Constanza; May Del Pozo, C.; Martín del Campo, Ana Victoria Data Economics and Artificial Intelligence in Latin America in this volume.

set forth in the consumer's law currently enforced in Chile, which indicates that it is obligatory²⁶ to keep the prices offered visible. Since in practice it is common to have a list price and to offer the consumer in person a lower price, this seems to clear the difficulty. There is a general price that can be known by all consumers and a special price offered in particular to some of them. Yet the question about why this consumer receives a special price while others do not remains unanswered. Is this discriminatory, pursuant to the provisions set forth in article 3 section c of Law 19496? Every time the drivers of the algorithm to set a different price for a certain consumer, and not for another, are correlations of data whose meaning is unknowable, we cannot answer that question, and it would be necessary to conduct a study on its tendency. In any case, the practice is a general one, and finding objective reasons for each case is difficult, so it is necessary to open the debate.

²⁶ L. 19496, Article 30: "Suppliers shall make known to the public the prices of the goods they sell or the services they offer, except for those which, because of their characteristics, must be regulated conventionally. The price must be indicated in a clearly visible way that allows the consumer to effectively exercise his right to choose, before formalizing or perfecting the act of consumption. Likewise, the tariffs of the establishments that provide services shall be stated. When goods are displayed in showcases, shelves, or racks, their respective prices shall be indicated there. The same information, in addition to the essential characteristics and benefits of the goods or services, shall be indicated on the Internet sites where the suppliers display the goods or services they offer and which comply with the conditions set forth in the regulations. The amount of the price must include the total value of the good or service, including the corresponding taxes. When the consumer cannot know by himself the price of the products that he wishes to acquire, the commercial establishments must maintain a list of their prices available to the public, in a permanent and visible way".

The case of financial services is even more complex, and it is a field where algorithms of this kind are commonly applied to make offers and establish credit conditions. Many data points could be taken into account to establish whether an individual is creditworthy or not. One of these is their record of breaches, creditworthiness, among others. But we do not know what the algorithm takes into account to determine the risk posed by an individual. But if it has been built on the basis of deep learning and open databases, it may be affected by discrimination criteria, like place of residence, naming features or other equally troublesome factors. As regards financial matters, article 3 section a of Law 19496 establishes that, if someone wants to contract a financial service but it is rejected, the consumer must "be informed by written notice about the reasons for the rejection, which should be based on objective conditions." It is worth wondering whether the determination of high risk by an algorithm is compelling enough, or if consumers should also be notified of the reasons for the model to establish that risk level in particular. Should the second option be taken, which seems more in line with the spirit of the legislation, it is hard to imagine how the decision of the algorithm could be

justified. In fact, credit models have been questioned, especially because their risk assessment and, therefore, the credit interest rate, is certainly quite dark for the consumer²⁷.

It is important to point out that in a society in which social mobility has historically been low and poverty, together with the lack of creditworthiness, has been historically linked to ethnic components, as in our society, it is possible for a model to take factors correlated with poverty such as name²⁸, place of birth, primary education institution, among others, to build a model that assigns a higher risk to people with such records and a lower risk to people whose factors are usually correlated with success²⁹. In short, an algorithm could easily take factors that have historically been the basis for social prejudice and boost them to an utterly unacceptable degree³⁰. Programmers and designers of the model can be totally unaware of this. They may even be looking for exactly the opposite, like fostering a more egalitarian, less segregated society by using mathematical designs. However, the agent resulting from their efforts may tend to social immobility, a re-ethnification of poverty and a further tightening of better-off social groups. This has been called “Weapons of Math Destruction”³¹.

On occasions it has been pointed out that, although algorithms can contain different kinds of biases, just like humans, these biases will always be fewer than those applied by people (Casey and Niblett, 2016, 437) so it is claimed that it is important not to inflate this risk. We do not agree with this since as far as human beings are concerned, we can ask them for their drivers. We may also force them legally to express them. Yet we cannot do the same with artificial agents, which are just unconscious entities that mechanically manipulate symbols to which they cannot bestow any meaning.

If due to a bias it is risky to use artificial agents in the private sphere, in the public domain it is even harder. Multiple efforts have been made to use artificial agents in order to automatize government actions. Thus, these models could perform actions that have been usually performed by humans, like the tasks made by social assistants or judges. If these mechanisms have a bias from data or from the objective function design, the result will

²⁷ Frank Pasquale (2015, 4) notes that: “A bad credit score may cost a borrower hundreds of thousands of dollars, but he will never understand exactly how it was calculated.” Véase también: Marron (2007, 111).

²⁸ Indeed, there have been cases where bias has been detected in Google’s result prediction algorithm. See: Pasquale (2015, 40).

²⁹ See: Ramirez (2019).

³⁰ The situation has been described in the following terms: “Mounting evidence shows that automated decision-making systems are disproportionately harmful to the most vulnerable and the least powerful, who have little ability to intervene in them—from misrepresentation to prison sentencing to accessing credit and other life-impacting formulas”. Noble (2018, 49).

³¹ Cathy O’Neil (2016, 3) coined the term, and describes the problem in dark colors: “Nevertheless, many of these models encoded human prejudice, misunderstanding, and bias into the software systems that increasingly managed our lives. Like gods, these mathematical models were opaque, their workings invisible to all but the highest priests in their domain: mathematicians and computer scientists... they tended to punish the poor and the oppressed in our society, while making the rich richer”.

be an amplified discrimination against the most vulnerable members of our society³². The act of discrimination is here spread automatically by the state itself, which is in theory the entity in charge of fighting against it.

An example that we should pay attention to comes from an algorithm for sentencing prediction designed some time ago, which the renowned Lex ex Machina research team, then based at Stanford University, made some experiments with (Surdeanu et al., 2019, 116-120). This was one of the first models using machine learning to predict the results from lawsuits related to brands and intellectual property. Once the analysis of its correlations and data points was conducted, it was found that the most relevant factors to predict the result were the identity of the judge and the name of the team of attorneys for the parties (Ashley, 2017, 124). The result is interesting because it shows the value of having a good judge and a solvent team of lawyers. But if the idea is to turn this algorithm into a non-predictive, sentencing system, cases will not be decided in terms of merits but on the identity of the lawyers, which would be evidently illegal.

When the risk posed by the offender to society is assessed, it is interesting that it seems to point at the personal traits, and not at the acts they may have committed. The risk assertion conducted by the model is based on who the person is, who their friends are, how they have been brought up and what their socioeconomic level is, rather than on the actions they are charged for and their previous behavior. It seems to be along the lines of the Offender-based Criminal Law, judging people on who they are and not on what they do -not appropriate in a democratic society. If it is decided the risk of crimes can be certainly predicted, it should not come as a surprise that a preventative system will be implemented against those suspected of being able to commit crimes in the future. Incredible as it may seem, this is exactly what has happened (O'Neil, 2016, 102). Robert McDaniel was visited by the police in 2013 without ever having committed a crime, because the Chicago City Police Department decided to carry out an algorithmic crime prevention program and he was selected as one of the four hundred people most likely to commit one in the near future³³. In China, in fact, there is already a policy in place to prevent future offenders³⁴. When, based on certain algorithmically evaluated records, the high possibility of future offending by a citizen is determined, the citizen is arrested and sent to a "re-education" camp.

³² Virginia Eubanks (2018, 11) describes the situation in the United States in the following terms: "Across the country, poor and working-class people are targeted by new tools of digital poverty management and face life-threatening consequences as a result. Automated eligibility systems discourage them from claiming public resources that they need to survive and thrive. In addition to the information provided by the public, businesses must keep a list of their prices available to the public, permanently and visibly."

³³ See the press report at <http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist> Visited on October 21st 2019.

³⁴ See press information about this in *The Washington Times* available at: <https://www.washingtontimes.com/news/2019/jun/25/china-s-chilling-pre-crime-prison-in-doctrina-tion-sy/> Visited on October 21st 2019.

The most important point is that judges, like the other organs of the State, must submit their actions to the law and the Constitution (Article 6 CPR), so that their actions must be framed within the legal norms that our legal system establishes, in the manner that those precepts set forth. This implies that in each specific case they must review whether or not the situation is adapted to the legal type and, once this is verified, they will apply the sanction or legal consequence pre-established by the law to the fact evaluated. If they act guided by arbitrary correlations, their acts will reflect these correlations and will not constitute, in themselves, applications of standards. Deep learning algorithms do not apply standards, in fact, connectivist type mechanisms are not capable of understanding them or manipulating symbols in order to make a subsumption, so they will always be, essentially, legal. In this sense, by making use of them, the application of rules and, therefore, the law as such, is renounced. This is not exclusive to judges, but is common to the entire State administration inasmuch as the executive branch must apply the law in force. The existence of a legal system implies the application of rules, and a mere mathematical correlation is not such. The question is where do we want to live as a society, whether in a place where the rule of

law prevails or where it does not. Since antiquity Aristotle (3.8-19, 1286a.) defended the rule of law against even the best of men. The rule of law is the foundation of modern democracy and, in its essence, freedom consists in being able to do what the law allows,³⁵ so that to renounce a normative system that regulates the conduct of individuals in society implies a departure from the concept of freedom that has served as the basis of our legal-political tradition.

There is one final issue related to the bias: how should it be corrected. In principle, once an algorithm has been detected to contain a bias, or even, when designing it and establishing its objective function, factors can be introduced to mitigate it or even to clear it out. For instance, if the algorithm for selecting CVs systematically discards those belonging to women, the tendency towards a balanced selection of men and women can be incorporated into the objective function. The main problem with this solution is that it is necessary to identify the biases our database may have, and to be careful when designing the objective function whenever it should be done consciously. This entails two risks: the first is the potential loss of efficacy in the system, and the second is the selection of the set of values to be incorporated.

³⁵ *Libertas in legibus consistit. Cic. De l. Agr. 2.102.*



There is one final issue related to the bias: how should it be corrected. In principle, once an algorithm has been detected to contain a bias, or even, when designing it and establishing its objective function, factors can be introduced to mitigate it or even to clear it out. For instance, if the algorithm for selecting CVs systematically discards those belonging to women, the tendency towards a balanced selection of men and women can be incorporated into the objective function. The main problem with this solution is that it is necessary to identify the biases our database may have, and to be careful when designing the objective function whenever it should be done consciously. This entails two risks: the first is the potential loss of efficacy in the system, and the second is the selection of the set of values to be incorporated.

Lastly, care should also be taken when tackling bias correction. In effect, when a bias is detected, it is often possible to eliminate it –not always, though. To do so, the objective function of the algorithm must usually be examined. What is it aimed at? Is the aim to hire employees with similar characteristics to those already working in the company (in which case it will reproduce the biases contained in the staff), or to modify it in any way (making it more diverse, lowering the number of days off, etc.)? If the idea is to vary the database feeding the agent, it will be necessary to consider this and program this kind of elements into the objective function. But it is essential to be able to see them in order to correct them. If the CV must not show gender, but the background selection algorithm unexpectedly segregates women by gender using statistic substitutes, as in the case of Amazon, it is hard to correct the bias without an explicit indicator on the applicant's gender condition. Amazon had to stop using its algorithm simply because since the CVs did not have to mention gender, it was impossible to suitably correct it. Surprisingly, correcting the bias brings to light any inequalities there might be within a group of humans. In this sense, it is very honest exercise.

Now this problem poses an evident question as to what values should be incorporated in a model to correct the biases it generates. This is a political and moral decision. In an axiological constitutional system, where values are incorporated into the rules of the Constitution, at least for the public sector, the answer has to be the ethical system on which the Constitution is grounded, especially in its first title "Basis for Institutionality" and in article 19, on the rights of individuals. Yet with no clear hints, it is not easy to see how to proceed.

Regarding the private sector, the doubts are maybe more relevant. Can a company project its values to society, suppressing terms and functions incompatible with those? Should it be axiologically neutral?

Since at least 2017, Google has an explicit social justice policy in its algorithms in order to correct the biases that they might contain. This is

known as Machine Learning Fairness (MLF)³⁶. It expressly incorporates non-discrimination policies in algorithm building so as to provide a more balanced vision on society in the results from the searches done by users. The idea is that they should not only show the most popular, but also the beliefs and practices of more vulnerable groups and minorities. This policy has been questioned lately every time there is an imposition on people who do not necessarily share the same values of the company. It even alters the network reality (Murray, 2019, l. 2102 and ff.). However, we will not delve into this issue because a detailed discussion would require an extension we cannot afford here.

4 conclusions

Based on our analysis we may conclude that artificial intelligence systems based on neural networks show various risks when they are applied to the field of Law as predictive mechanisms.

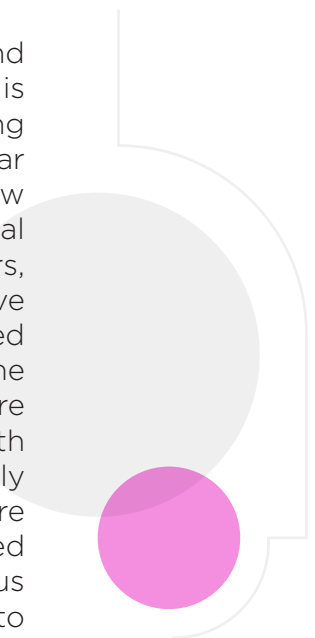
The so-called “bias problem” is known as “prejudice” in continental thinking, which points at the existence of a judgement previous to another one, which affects or modifies the second. It evokes a debate that encompasses a much greater theoretical and anthropological entity: the question whether human knowledge comes from evidence or if it is the result from a method, i.e., from a demonstration process. The answer from the classical, Roman and medieval world was that the starting point for knowledge was some kind of external reality, built mainly from itself, which appeared in human conscience as a phenomenon. In this context, the pre-judgement –i.e. the judgement prior to the formal thinking- not only was not a defect but it aimed at the very condition of human cognitive structure. It is Modernity with its new methodological turns and its intention to apply the conditions of “purity” from logical-empirical sciences to human knowledge that turns the prejudice into a (supposed) thinking error. Thus, in order to be pure, suitable or efficient, the reflection should be devoid of any contents previous to the encounter between the interpreter and the interpreted. In this context, the perfectionist desire to eliminate all preconceptions aims at reaching a decision through a pure methodology.

³⁶ See the document *Responsible AI Practices*, available at <https://ai.google/responsibilities/responsible-ai-practices/?category=fairness>
Consulted on October 22, 2019.

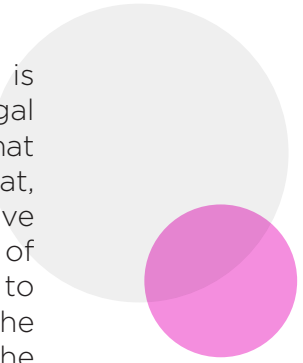
However, this is impossible because all human judgements are an action of consciousness, and consciousness is the basis for identity: nothing can separate consciousness from its preconceptions. This does not mean that the agent as an individual cannot or should not make an effort to separate the preconceptions that are not related to the formulated judgement. In this sense, all judgements and all human decisions are shaped by preconceptions and prejudices. The right way to act is not to suppress them, because it is not possible, but to regulate them with ethical and justice criteria.

From the point of view of artificial intelligence, there is one more problem to be considered: the difficulty for a mechanical discernment to take into account the contexts, thus generating propositions that do not properly interpret reality. The reason for this aporia lies in the fact that the way to think about computer structures does not have an end purpose but is linear instead, meaning that it cannot grasp the nature of purposes or the abstract nature of its contemplation. Artificial intelligence may know the purpose integrated by the programmer but cannot perform a contemplative, synthetic or ductile action of that end or the means to achieve it. In other words, the machine's reason does not seem to be speculative or practical, just purely logical, in a linear sense.

Although the most evident risk is reproducing and amplifying biases and arbitrary discriminations in the data with which the network is fed, this is not the only one. There is also an evident risk in the fact that deep learning models simply make correlations and determine results through linear analysis that are not fully compatible with the structure of Law. The Law has been designed essentially as a system of rules that regulates social situations by assigning (subjective) rights to different legal operators, depending on what they have been assigned by nature or by law (objective law). This model is basically incompatible with a system of data based correlations system every time their way of operation is different. The starting point of a model is a set of rules, from which subsumptions are made to reach concrete results, while the other one correlates data with results, without understanding or paying attention to the explicitly formulated rules. Although the old expert systems from the 80's were capable of understanding rules, they had such limitations that they tended to collapse when they were configured outside very specific areas, thus failing to reach the expected results. However, the current approach to neural networks implies an implicit waiver to the presence of rules, so their operation is in principle incompatible with a legal model based on them.



The solutions to this difficulty are not clear-cut or simple. Although it is true that a well-trained algorithm could predictively determine legal solutions similar to those obtained through a system of rules, the fact that the results obtained are not determined by said rules implies that, notwithstanding the good result, the model does not apply the objective Law to solve the case, which is a very relevant problem from the point of view of Law operators. In some sense, this issue shows similar difficulties to those suggested by Kantorowicz (1906, 10-17) when he posited the irrelevance of dogma when solving a conflict, although he left it to the judge to decide on the real resolution and not to a correlations-based algorithm.



Referencias bibliográficas

Ashley, D. K., *Artificial Intelligence and Legal Analytics. New Tools for Law Practice in the Digital Age*, Cambridge University Press, Cambridge, 2017.
Aristóteles, *Política*, 3.8-19.

Barcia, R., *Diccionario Etimológico de la Lengua Castellana*, José María Faquinetto Editor, Madrid, 1889, voz “sesgo”.

Broussard, M., *Artificial Unintelligence. How Computers Misunderstand the World*, MIT Press-kindle, Cambridge MA-Londres, 2018.

Buolamwini, J. y Gebru, T., Gender Shades: Intersectional Accuracy Disparities, en *Commercial Gender Classification in Proceedings of Machine Learning Research*, 81, 2018. Disponible en: http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline
Consultado el 12 de junio de 2019.

Casey, A. J. y Niblett, A., Self-driving Laws, en *University of Toronto Law Journal*, 66-4, 2016, pp. 429-442.

Chopra, S. y White, L. F., *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press-Kindle, Michigan, 2011.

Eubanks, V., *Automating Inequality. How High-Tech Tools profile, police, and punish the poor*, St. Martin's Press, Nueva York, 2018.

Gurney, K., *An Introduction to Neural Networks*, UCL Press Limited, Londres, 1997.

Hinton, G., Salakhutdinov, R., Reducing the Dimensionality of Data with Neural Networks, en *Science*, 313, 2006.

Huang, X., Baker, J. y Reddy, R., A Historical Perspective of Speech Recognition, en *Communications of the ACM*, 57-1, 2013, pp. 93-103.

Kai-Fu, L., *AI Superpowers: China, Silicon Valley and the New World Order*, Houghton Mifflin Hancourt, Boston, 2018.

Kantorowicz, H., *Der Kampf um der Rechtswissenschaft*, Carl Winter, Heidelberg, 1906.

Lee, K. F., *AI Super-Powers. China, Silicon Valley, and the New World Order*, Houghton Mifflin Harcourt, Boston-Nueva York, 2018.

Marron, D., Lending by Numbers': Credit Scoring and the Constitution of Risk within American Consumer Credit, en *Economics and Society*, 35, 2007.

Murray, D., *The Madness of Crowds. Gender, Race and Identity, Bloomsbury Continuum-Kindle Edition*, Londres-Oxford-Nueva York- Nueva Delhi-Sidney, 2019.

Noble, S. U., *Algorithms of Oppression. How Search Engines Reinforce Racism*, New Yor University Press-Kindle, Nueva York, 2018.

O'Neil, C., *Weapons of math Destruction. How Big Data Increases Inequality and Threatens Democracy*, Crown Publishers, Nueva York, 2016.

Pasquale, F., *The Black Box Society. The Secret Algorithms that Control Money and Information*, Harvard University Press, Cambridge MA-Londres, 2015.

Pearl, J. y Mackenzie, D., *The Book of Why. The New Science of Cause and Effect*, Basic Books, Nueva York, 2018.

Radford, A., Metz, L. y Chintala, S., *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks in ICLR*, 2016. Disponible en:
<https://arxiv.org/pdf/1511.06434.pdf>
Consultado el 8 de agosto de 2019.

Ramirez, E., *Privacy Challenges in the Era of Big Data: The View from the Lifeguard's Chair*. Disponible en:
https://www.ftc.gov/sites/default/files/documents/public_statements/privacy-challenges-big-data-view-lifeguard's-chair/130819bigdataaspen.pdf
Consultado el 15 de Octubre de 2019.

Roget, P. M., y Davidson, G. W., *Roget's Thesaurus of English Words and Phrases*, voz "bias", Penguin, Londres, 2003.

Rumelhart, D., Hinton, G. y Williams, R., Learning Representations by Back-Propagating Errors, en *Nature*, 323, 1986, pp. 533-536.

Surdeanu, M., Nallapi, R, Gregory, G., Walker, J. y Manning, C., *Risk Analysis for Intellectual Property Litigation in Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ACM, Nueva York, 2019, pp. 116-120. Disponible en: <https://nlp.stanford.edu/pubs/icail11.pdf>
Consultado el 21 de Octubre de 2019.

Surden, H., Machine Learning and Law, en *Washington Law Review*, 89, 1, 2014.

Unión Europea, *High Level Expert Group on Artificial Intelligence, Ethic Guidelines for Trustworthy AI*. Disponible en: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
Consultado el 13 de Junio de 2019.

Zhang, M., Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software, en *Forbes*, 1 de julio de 2015. Disponible en: <https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/#1530ee48713d>
Consultado el 12 de junio de 2019.

Zuiderveen Borgesius, F., *Discrimination, Artificial Intelligence and Algorithmic Decision-Making (Directorate General of Democracy, 2018, Estrasburgo)*. Informe para el Consejo de Europa. Disponible en: <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73>
Consultado el 10 de Octubre de 2019.

Disclaimer. The opinions expressed in this publication are those of the authors. They do not purport to reflect the opinions or views of the CETyS or of any other organization involved in the project.