Oñati Socio-Legal Series (ISSN: 2079-5971)
Oñati International Institute for the Sociology of Law
Avenida Universidad, 8 – Apdo. 28
20560 Oñati – Gipuzkoa – Spain
Tel. (+34) 943 783064 / opo@iisj.net / https://opo.iisj.net

Biases and debiasing in human and artificial intelligence

OÑATI SOCIO-LEGAL SERIES FORTHCOMING
DOI LINK: HTTPS://DOI.ORG/10.35295/OSLS.IISL.2304
RECEIVED 21 MARCH 2025, ACCEPTED 17 JULY 2025, FIRST-ONLINE PUBLISHED 3 SEPTEMBER 2025

PATRIZIA CATELLANI*
MARCO PIASTRA*

Abstract

When we make decisions or argue, cognitive, emotional, and motivational factors often lead us to use mental shortcuts. These can speed up reasoning but can also lead to systematic biases. To counteract them, various debiasing strategies can be tested experimentally. In the justice system, it is crucial that the decision-making procedures match the capabilities and limits of human cognition. Artificial intelligence is also affected by biases, but these are not due to the subjectivity of the algorithms, but to flaws in the formal methods. Rigorous methodological analysis is essential to improve awareness and control. Algorithms can be adjusted and monitored to ensure that their predictions and actions reflect reality and serve their intended goals. Whether in human or artificial intelligence, debiasing strategies help to reduce bias and improve the quality of reasoning and decisions.

Key words

Bias; debiasing; causal reasoning; responsibility attribution; artificial intelligence

Resumen

Cuando tomamos decisiones o discutimos, los factores cognitivos, emocionales y motivacionales a menudo nos llevan a utilizar atajos mentales. Estos pueden acelerar el razonamiento, pero también pueden dar lugar a sesgos sistemáticos. Para contrarrestarlos, se pueden probar experimentalmente diversas estrategias de eliminación de sesgos. En el sistema judicial, es fundamental que los procedimientos de toma de decisiones se ajusten a las capacidades y límites de la cognición humana. La inteligencia artificial también se ve afectada por sesgos, pero estos no se deben a la subjetividad de los algoritmos, sino a fallos en los métodos formales. El análisis

Funding institutions: This article is part of the results of the research project "Communicating Science. Mediation and Mediators of Scientific Knowledge in a Complex Society"," which was funded by the Catholic University of the Sacred Heart (funding line D.3.2/2022).

* Corresponding author: Patrizia Catellani. Catholic University of the Sacred Heart, , Largo Gemelli, 1, 20123, Milan, Italy. Email: patrizia.catellani@unicatt.it ORCID: https://orcid.org/0000-0002-7195-8967

* Marco Piastra. University of Pavia, Italy. ORCID: https://orcid.org/0000-0003-2556-5254

metodológico riguroso es esencial para mejorar la concienciación y el control. Los algoritmos pueden ajustarse y supervisarse para garantizar que sus predicciones y acciones reflejen la realidad y sirvan a los objetivos previstos. Tanto en la inteligencia humana como en la artificial, las estrategias para eliminar sesgos ayudan a reducir los sesgos y a mejorar la calidad del razonamiento y las decisiones.

Palabras clave

Sesgo; eliminación de sesgo; razonamiento causal; atribución de responsabilidad; inteligencia artificial

Table of contents

1. Introduction

In causal reasoning in everyday life as well as in legal reasoning, we can use mental shortcuts or heuristics. They can be very useful strategies in problem solving and decision making, allowing us to take into account our previous experiences to optimize reasoning (Gigerenzer 2008). Thus, heuristics can help us make correct decisions and reason well in both factual and legal matters. However, sometimes they are overused, which can increase the likelihood of biases, that is, systematic distortions in our reasoning process (Dale 2015). This can be particularly the case in context-specific situations (e.g., chronic illness) where emotional or physical states such as pain or fatigue influence cognition independently of heuristic reasoning (Savioni and Triberti 2020).

Understanding the reasons for incurring in biases is important to develop strategies to mitigate them and thus improve the quality of reasoning. Something similar is happening in the field of artificial intelligence. Algorithms or the data they are applied to can be biased for various reasons. But even in this case, it is possible (and more necessary than ever) to develop strategies that allow us to reduce the probability of bias or minimize the likelihood of its occurrence.

In this paper, we look at biases and their mitigation in human and artificial intelligence. We will start with the distinction between surface and deep information processing and then focus on some heuristics or cognitive shortcuts, so-called biases, that we often use in thought processes. In particular, we will focus on the biases that can occur when thinking about cause-effect relationships. We will also see how we can deal with our cognitive limitations by applying strategies to improve the functionality of our thinking processes. We will then turn our attention to artificial intelligence. By discussing how algorithms can exhibit bias despite the absence of ‘unpredictable subjectivity’, we will highlight the challenges of detecting biases in formal and deterministic methods. We will then show that even in the field of artificial intelligence, a thorough analysis of causal relationships can be of great benefit to develop fair and unbiased automatic decision-making systems.

2. Heuristic versus systematic processing

In his famous book *Thinking, Fast and Slow* (2011), Daniel Kahneman states that each of us has two different ways of thinking: a more superficial and heuristic way (called *System 1*) and a deeper and systematic way (called *System 2*). Depending on the circumstances, we tend to favour one way of thinking over the other.

For example, let us say I come home from work on a winter evening. It’s not late yet, but it’s already quite dark and my house is on a secluded street with not much going on. As I approach the gate of my house, I see a tall figure with a hood over his head, who is obviously male, walking towards me. What do I do at this moment? I quicken my pace, quickly insert the key into the gate to open it and then close it behind me as quickly as possible. But this person is now very close to me and says, “Mom, what are you doing, leaving me outside?” At that moment, of course, I realize that it is my son, a one-meter-ninety-two tall young man whom I had not recognized from a distance.

Have I made a mistake in my thinking in this situation? Obviously yes, in the sense that I applied a superficial and heuristic way of thinking and, based on some clues and my previous experience with potentially dangerous situations, quickly came to a conclusion that was actually wrong. But in another sense, I did not make a mistake, because there are situations where heuristic thinking works better than systematic thinking. Typically, we resort to this type of thinking when we have little time or information available or when we are strongly emotionally involved. In my case, the strong emotion I felt was obviously fear, and the situation certainly called for a quick decision, a decision that turned out to be wrong but would have been right if it had been a malicious person. Superficial information processing, which occurs quickly and is based on previous experience rather than on the analysis of the current situation, is therefore more prone to error than in-depth processing. However, it often proves to be appropriate and useful because it allows us to reach a decision quickly with limited mental energy.

Take the case of experts in a particular field who are confronted with familiar situations for which they have developed very effective solution routines. In these cases, paradoxically, relying on established routines and thus on intuitive and quick thinking may prove to be the best option. This is the viewpoint of German psychologist Gerd Gigerenzer, who explores this topic in the book *Gut Feelings. Short Cuts to Better Decision Making* (2008), where 'gut feelings' can literally be translated as 'gut instinct'. It is not for nothing that it is sometimes said that the gut is our second brain. In some cases, this does indeed seem to be the case. Gigerenzer has conducted several studies in which he examined the thinking of doctors with a high level of expertise. He has found that these doctors can often make the correct diagnosis before the results of further tests are available or even independently of them. These doctors therefore proceed intuitively, based on their previous experience, and of course they too can make mistakes. However, if the routines they have acquired are valid, it is very likely that this error will not occur. This also applies to experts in other fields of knowledge.

Gigerenzer speaks of ecological rationality: the expert uses his mental energies sparingly and devotes his deep reasoning mainly to new and complex cases, while applying automatic routines in other cases. In fact, the expert's greatest skill may lie precisely in knowing when it is worth going deeper when the situation is different from the past. In this case, a true expert demonstrates the ability to abandon their routines and automatic processes and fearlessly engage in deeper analysis and, if necessary, change.

3. Selection of the cause

As we have seen, heuristic reasoning has its advantages, but problems arise when this reasoning is applied even though the circumstances would require deeper reasoning. In this case, we can easily fall into genuine fallacies or biases, i.e. systematic distortions in our assessment of reality. Some of the most common biases concern causal explanations of events and the attribution of responsibility for these events.

First, although an event often has multiple causes, we tend to select only those that are most accessible to us for various reasons, such as simpler rather than complex causes, or coherent rather than incoherent causes that are compatible with our worldview. Similarly, we generally tend to select causes that are close in time and give them more importance than causes that are far away. In the case of flood damage, for example, we tend to identify the fact that the authorities did not warn citizens in time (*proximate cause*) rather than hydrogeological instability due to inadequate maintenance (*remote cause*) as the cause. Furthermore, when there are several possible causes for a negative event, we tend to consider *human causes* more important than *natural causes*. For example, in one experiment participants read different versions of negative events. In all versions, two causes were presented, one human and one natural, which were presented as proximate or remote causes depending on the version (McClure *et al.* 2007). One of the proposed events was a fire, which was described in one version as follows:

"A boy walking in the woods stops and lights some undergrowth on fire (*remote human cause*); the wind fans the flames and causes a fire (*proximate natural cause*)." In another version, the fire was described as follows: "A shard of glass left in the forest focuses the sun's rays on the undergrowth, causing it to catch fire (*remote natural cause*). A man, unaware of the fire's origin, stops to stoke it, causing a fire (*proximate human cause*)."

Regardless of whether the cause was presented as proximate or remote, participants in this study considered the human cause to be more important than the natural cause.

In the field of human causes, we tend to make further distinctions. One of these is the distinction between *controllable and uncontrollable causes* by the persons involved in the event. Even within controllable causes, we make a further distinction that affects the perceived relevance of the cause itself, namely whether the cause was *intentional or not*. An example from research can help to illustrate how the perception of intentionality influences the evaluation of event causes (Martin and Cushman 2016). The scenario is that of a doctor treating a patient with a hearing problem who has two different treatments available, one with a success rate of 66 and the other with a success rate of only 33. To make the best choice, the doctor should opt for the first treatment, i.e. the one with the higher success rate. Depending on which experimental condition they were assigned to, participants were faced with one of two different versions of this scenario at this point. In the first version, the doctor chooses the first treatment, i.e. the one with the higher success rate, but the patient unfortunately suffers permanent hearing loss (*intentional condition*). In the second version, it turns out instead that the patient has an allergic reaction to the drug used in the first treatment, so the doctor is forced to administer the second treatment, which has a lower success rate (*unintentional condition*). In this version, the patient also suffers permanent hearing damage. Participants are then asked to rate the causal role of the doctor in relation to the harm suffered by the patient, and there is a clear tendency to attribute a more important causal role to the doctor who had the choice (*intentional condition*). This doctor did indeed behave as he should have done by prescribing the treatment with the greater chance of success. However, in the event of an unfortunate outcome, this doctor was ascribed a greater causal role than the doctor who simply prescribed the only tolerable treatment for the patient (*unintentional condition*). The very idea that the first doctor had more options than the second caused the participants in this study to reinforce the causal role of the first doctor's actions, even if it was an intrinsically correct action.

In summary, when we think about events, especially negative events, we tend to choose between the different possible causes and 'favour' some of them over others. We favour the causes that are most easily accessible to our minds, i.e. human causes over natural ones, and within human causes we pay more attention to controllable and intentional causes. In short, we consider humans to be the primary cause of negative events and tend to overestimate the ability of those involved in an event to control it and act intentionally.

4. Attribution of responsibility

When we evaluate negative events, we not only think about the causes, but also instinctively tend to attribute responsibility and blame (Roese and Olson 1996). Even in doing so, systematic biases sometimes occur. For example, when we attribute responsibility and blame not only based on evaluating the information we have about the event and the people involved, but also based on other information about these people that has nothing to do with the event.

In some cases, the information outside the event that is taken into account in the evaluation relates, for example, to the victim of the event. This is shown by studies in which participants read the description of a case of a doctor treating a patient with a gunshot wound (Alicke *et al*. 2008). Depending on which experimental condition the participants were assigned to, the patient was presented either as: a) a construction worker who was shot when he tried to intervene in an argument between a woman and her armed husband (*positive* presentation of the patient); b) a former convict who was shot by a police officer when he tried to draw his gun during a check following a complaint against him (*negative* presentation of the patient). In both versions of the case, the patient was said to have died of a haemorrhage that had occurred during an operation. Thereafter, the versions presented to the participants differed again depending on the experimental condition: a) the autopsy revealed that the patient could have been saved if the doctor had diagnosed a blood clotting disorder the patient was suffering from in time (*controllable outcome*); b) the autopsy revealed that the patient had already lost too much blood by the time he arrived at the hospital to be saved (*uncontrollable outcome*). The results of the study showed that, as usual, greater responsibility was attributed to the physician when the outcome was presented as controllable rather than uncontrollable. However, and this is the most interesting finding of the study, this responsibility was significantly higher when the victim was portrayed positively (i.e. as a construction worker) rather than negatively (i.e. as an ex-convict).

To summarize, the study confirms our automatic tendency to focus our attention on information about the main characters of an event that we process for various reasons, but which should not be considered relevant for attributing responsibility for the evaluated event.

5. The influence of psychophysiological factors and their management

Distortions about the causes and responsibilities of events are just some of the mistakes we can make in legal reasoning. In this and the following sections, we will look at other possible biases, but we will mainly see how they can be reduced by applying appropriate mitigation strategies.

As mentioned earlier, an important source of bias lies in the structure of our cognitive system, which has high speed and power in heuristic information processing, but limited ability to process information deeply and systematically. Our processing capacity is also limited in terms of maintaining a constant level of performance over time due to many internal (psychophysiological, emotional, motivational) and external (physical or social) factors.

An example of the influence of psychophysiological factors are the results of a study conducted in Israel in which the decisions of judges assessing the conditions of prisoners' parole were analysed (Danziger *et al*. 2011). The judges examined the cases one after the other, day by day, so that it was possible to compare the decisions at different times of day and on different days of the week. By analyzing more than 1000 files, the researchers discovered a regular pattern in the favour or disfavour of the decisions. The least favourable decisions were made in the hours immediately before the lunch break (judges' hunger state), while the most favourable decisions were made in the hours immediately after the break (judges' satiety state). For this reason, the researchers called the observed effect the 'breakfast effect'. Of course, the link with hunger and satiety is only one interpretation of this correlation, which is entirely consistent with what we know about the effects of different physiological states on cognitive performance, emotional state and decision-making. In fact, there are numerous other studies that have investigated the effects of not only hunger, but also fatigue, mood and other psychophysiological factors on decision-making (Gawronski *et al*. 2018).

What research on the breakfast effect has not been able to determine, due to the limited and correlative nature of the data used, is the exact mechanism linking hunger (or satiety) to judges being less or more lenient in their decisions. The main hypothesis regarding this mechanism is that people are partly guided in their decisions by cues from self-perception (Damasio 1996). They rely not only on external elements (what we read or hear about a case) but also on some information about their internal state when processing information and formulating judgments. These cues provide continuous feedback about the sensations and emotions that a situation, task or object of judgement triggers. This is usually a functional process for correctly evaluating situations and making decisions. In some cases, however, these psychophysiological cues can mislead us by associating negative or positive states, regardless of what we are doing in the current activity.

How can we deal with these psychophysiological disturbances so that they do not affect the quality of our evaluations and decisions? One way is to strategically plan our activities and take into account the times when we are more prone to these types of errors. Another option is to make some tasks easier, possibly the less pleasant ones, so that they are more automatic and we do not waste our valuable cognitive resources. Conversely, we can try to focus our efforts on more complex, but also more intellectually and professionally satisfying tasks in order to make positive use of the psychophysiological feedback that results from performing such activities. Some studies (Inzlicht *et al*. 2014) suggest that when we perform a task independently and with commitment, we tend to complete it with more effort and attention and are less prone to errors, distractions and cognitive distortions.

6. Dealing with metacognitive limits

Another important source of thinking errors are our metacognitive limitations. Metacognition is an advanced mental function that allows us to be aware of the content of our thoughts and to check the flow of our reasoning (more precisely, our cognition) by temporarily taking an external point of view (Metcalfe and Shimamura 1994). This ability is necessary to recognize possible errors in thinking, not only in retrospect, but also while we are performing an evaluation task or forming a judgment. Once we have identified a possible bias, we can return to the process that led or is leading us in a particular direction to analyse it more thoroughly and possibly correct it.

One way to activate metacognition is to check our knowledge, reasoning and intuitions by taking a different, independent and more critical point of view. This does not necessarily lead us to understand what is wrong with a potentially biased argument, but it does lead to a closer examination of the steps we take in processing information.

For example, it can be useful to take time at the end of information processing to consider the diametrically opposite position to the one we have reached. By trying in this way to find elements that support a conclusion other than the one already reached, we can regain control of our evaluation processes and free them from the confirmation trap, i.e. the tendency to look only for elements that support an opinion already formed. A similar technique is the so-called devil's advocate, which was originally developed for group decision-making (MacDougall and Baum 1997), but can also be applied to individual decisions. Rather than trying to support another position, in this case it is about questioning our position and pointing out possible inconsistencies, ambiguities or weaknesses. Questioning our judgment and reasoning may not be pleasant, but it can also be an interesting challenge that encourages us to carefully, accurately and rigorously analyse what we may have previously taken for granted, driven by the motivation to reach a quick and certain conclusion.

7. Restructuring of decision-making processes

A specific source of bias in judicial decision making arises from the potential mismatch between the formalized procedures for performing certain tasks, such as evaluating evidence or assessing various aspects of a case, and the way our minds tend to perform these tasks. This is the case, for example, when rules and procedures require us to perform mental operations that are formal, logical and legally correct, but at the same time counterintuitive and far removed from our spontaneous behaviour.

For example, when we retrospectively evaluate behaviours that led to unforeseen or unusual consequences, we are asked to judge what the person knew, thought or intended at the time of the event, deliberately ignoring what we know in hindsight about how that knowledge, beliefs and intentions led to an actual outcome. Similarly, there are situations where certain evidence gathered during the investigation must be ignored during the trial. Again, the judge is forced to evaluate a situation without considering information that is known to him.

In such cases, the judge must simulate his own judgment in the presence or absence of certain elements: a task to which those working in the judiciary are quite accustomed (Catellani 2010), but which tests both the limits of our information processing capacity and our metacognitive monitoring capacity.

How can we deal with this situation? Some experimental studies have provided interesting insights into informal strategies and procedures that can be applied to decision-making processes to improve their quality. For example, the technique of so-called bifurcation, i.e. the division of tasks within the panel of judges (a technique that is therefore not applicable to single-judge trials), has been tested in retrospective judgments about past events and in how to prevent harm assessments from interfering with judgements of voluntariness and responsibility (Smith and Greene 2005). By asking groups of simulated jurors to assess the same case together or by forming separate subgroups to assess harm and responsibility, it was found that in the latter case the interference between the two assessments was significantly reduced.

Research in the field of organizational psychology and decision-making processes also offers numerous insights that could be used to improve the flow and quality of work in offices, law firms and courtrooms. For example, research in the field of nudging (Hertwig and Grüne-Yanoff 2017) shows that with sufficient knowledge of the most functional and problematic aspects of an individual or group decision-making process, it is possible to favour more functional procedures by making them more accessible while maintaining the necessary flexibility when circumstances require the use of other procedures.

To summarize, legal reasoning, like any other form of reasoning, is subject to possible errors (biases) due to an erroneous recourse to one or the other of the two types of information processing, i.e. heuristic rather than systematic. These errors can be caused by various factors: limited availability of resources for thorough processing; the tendency to fall back on the heuristic processing system and the inability to notice this deviation and get the reasoning back on track; difficulties in applying external rules and criteria that do not fit well with the way our cognitive system is used to functioning.

Depending on the type of error we fall into and the source from which it originates, we can apply different mitigation strategies to compensate for some distortions, prevent others or, more simply, become aware of their existence. With this heightened awareness and willingness to question ourselves and review habitual ways of doing things, it is possible to create conditions to develop more effective thought processes for the activities we are asked to perform.

8. When algorithms enter the scene

Any application of an algorithm to a particular input with the aim of producing a particular output corresponds to an inferential process. The fact that such a process is automated and controlled by a computer does not change its nature, even if the implemented process is the result of an indirect method of learning from data. The increasing prevalence of methods that make use of artificial intelligence and machine learning has led to more attention being paid to the problem. The reason for this is, on the one hand, the obvious expansion of the possibilities for applying artificial intelligence to delicate and quite sensitive situations and, on the other, the fact that the great technical complexity of the systems used can give rise to fears of a loss of control. In descriptions of automatic reasoning processes carried out by algorithms and AI, a certain subjectivity is often attributed to these systems, as if they had an autonomy that goes beyond the mere reproduction of certain thought functions in a formalized and precise manner. However, algorithms execute deterministic, predefined and fully repeatable procedures, unless random behaviours are intentionally introduced to simulate a deceptive form of subjectivity. In the following sections, we will explore how bias can manifest itself in automated formal reasoning processes and the importance of a deeper understanding to recognize and correct such processes.

9. An example of algorithmic bias

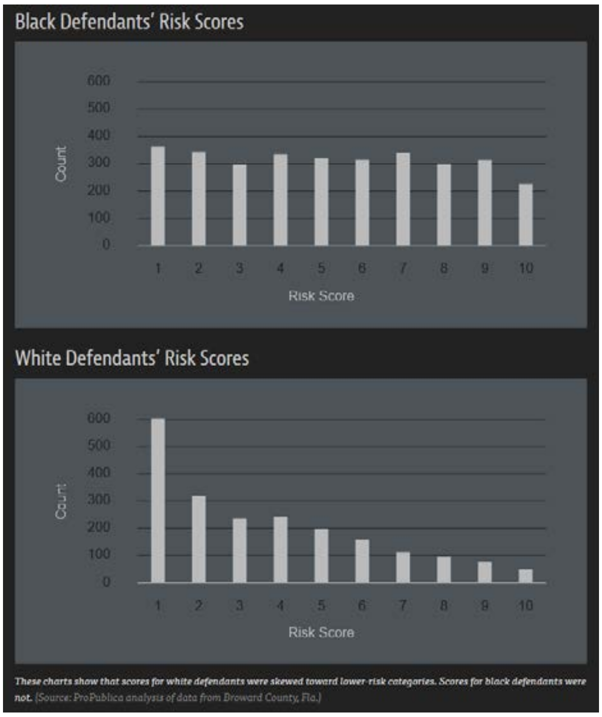
Often, descriptions of the functions of algorithms and artificial intelligence (currently the most advanced form of algorithms) tend to attribute a degree of subjectivity to these systems. It is as if they possess a form of autonomy that is clearly distinct from the capacity for automation that can occur under certain conditions. However, algorithms perform deterministic, predefined, and completely repeatable functions. This is true unless there is manipulation or an explicit intention to simulate a deceptive form of subjectivity by introducing random behaviours (which are more apparent than real).

Since 2008, some judicial systems in the United States have used an algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). This algorithm has been at the centre of extensive debate about the risks it poses in sensitive decision-making processes (Kirkpatrick 2017). The intention behind the introduction of COMPAS was to streamline judicial activity in the criminal justice system, particularly for minor offenses and the assessment of the likelihood of reoffending.

COMPAS works based on the answers given by the person being assessed to a questionnaire developed by behavioural experts. The questionnaire contains 137 questions on the person's psychological characteristics, their general attitudes, and their relevant criminal history. Based on these answers, COMPAS generates a score that estimates the risk of recidivism. The actual methodology and validation criteria used by COMPAS are not disclosed, as the algorithm is proprietary and patented by Northpointe Inc. (now Equivant Inc.).

In 2016, the Wisconsin Supreme Court upheld the use of COMPAS in the judiciary and rejected an objection to its admissibility. However, the court recommended that caution and doubt be exercised when using the algorithm (Smith 2016). In the USA, the debate about the possible bias of COMPAS was triggered by a study conducted by ProPublica, an investigative journalism organization serving the public interest (Angwin and Larson 2015). According to the study published in 2016, which analysed data from 18,610 individuals in Broward County, Florida, the algorithm tends to overestimate recidivism risk for people of colour, while underestimating it for white individuals. ProPublica has made both a detailed description of their methodology and the results available online. At first glance, ProPublica's data is striking and clearly supports the suspicion of bias in COMPAS, as shown in Figure 1.

FIGURE 1



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

Figure 1. Statistical distribution of Risk Scores generated by the COMPAS algorithm for people of color and white individuals, according to ProPublica's statistical analysis.

As was to be expected, this triggered a lively debate characterized by an intensive exchange of opinions and scientific assessments. During the discussion, in addition to the initial concern about the bias of COMPAS, broader doubts were expressed about the methods used to identify bias in an algorithm. Specifically:

1. Based on the data, it is evident that COMPAS does not make significantly different errors than humans when assessing recidivism risk (Dressel and Farid 2016).
2. In attempting a more precise statistical formalization of what bias is, it becomes clear that the concept is much more difficult to define than one might think (Corbett-Davies *et al.* 2016).

Type 1 doubts arose when comparing the performance of the COMPAS algorithm with that of human judges. This problem, as we will explore later, has a logical basis in the design of algorithms, particularly those that use machine learning. Machine learning relies on available data that reflects human experience. If this data contains biases, it is difficult to eliminate them.

Type 2 doubts, on the other hand, point to several challenges that may seem purely quantitative and technical at first, but can also provide intuitive and more comprehensive insights. To explain this, let us apply a little formalism. Let A stand for a sensitive characteristic of a person, such as skin colour, gender, or other attributes. X stands for several other characteristics of the same person that are not necessarily sensitive. Y is the actual outcome we want to predict. In the case of COMPAS, whether the person will commit further crimes in the future. Finally, let $\hat{Y}$ be the prediction generated by the algorithm. In the case of COMPAS, the prediction depends on the assigned risk score: If the score exceeds a certain numerical threshold, the person is considered at risk.

10. Detecting and correcting algorithmic bias

One of the most important criteria for fair predictions, i.e. the absence of algorithmic bias, is the so-called *conditional demographic parity*. It can be expressed by the following formula:

$$\left\langle \ \tilde{Y} \perp A \,\middle|\, X \ \right\rangle$$

The formula states that the prediction $\tilde{Y}$ of the algorithm should be independent of the sensitive characteristic A for given characteristics X of the individual. In other words, for the same X, the algorithm should not produce different predictions $\tilde{Y}$ when A varies. This is certainly a desirable property, but if the algorithm is based on data reflecting actual experience, this must be achieved:

$$\langle \ Y \perp A \mid X \ \rangle$$

In the case of COMPAS, this would be true if the vast majority, if not all, of the actual reoffending outcomes were truly independent of variable A. In other words, and by extension, this condition is met if there is no bias in the actual data. As we can well imagine, this condition is rarely guaranteed in practice:

A second and different criterion for fairness is called *predictive parity*:

$$\langle \ Y \perp A \mid \tilde{Y} \ \rangle$$

In this case, it is desirable that the rate of correct predictions (i.e., where $\tilde{Y}$ matches Y) is independent of the sensitive property A when the predicted value $\tilde{Y}$ is given.

A third and additional criterion is called *predictive error parity*, which is in a sense the inverse property of the previous criterion:

$$\langle \ \tilde{Y} \perp A \mid Y \ \rangle$$

This means that, given the actual value Y, the error rate made by the algorithm should be independent of the sensitive property.

To remove possible ambiguities, especially regarding the practical distinction between the three criteria, it is helpful to analyze some of the statistics reported for COMPAS.

FIGURE 2

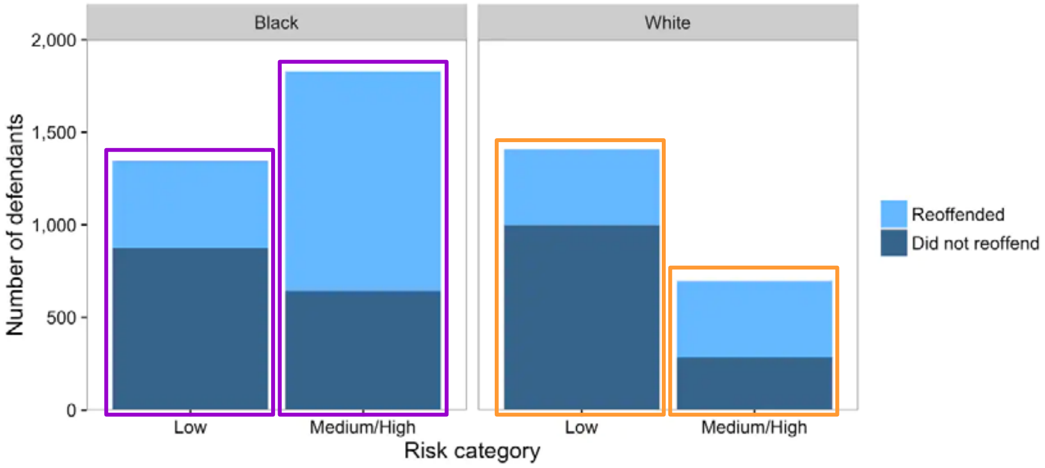


Figure 2. Evaluation of conditional demographic parity (Black vs. White) in the COMPAS algorithm.

In Figure 2, the columns represent the prediction $\hat{Y}$ of the algorithm. The height of each column is proportional to the number of people assigned a low or medium/high risk class by COMPAS. The two colours within each column indicate the actual outcome Y: whether the person relapsed (blue) or not (light blue). The figure is divided into two sections corresponding to the two values of the sensitive characteristic A: Black on the left and White on the right.

The first criterion, *conditional demographic parity*, would require the ratio between the heights of the Low and Medium/High columns be the same in both sections. However, this is not the case: The ratio between the heights of the two orange-coloured columns on the right is exactly the opposite for the two purple-coloured columns on the left.

FIGURE 3

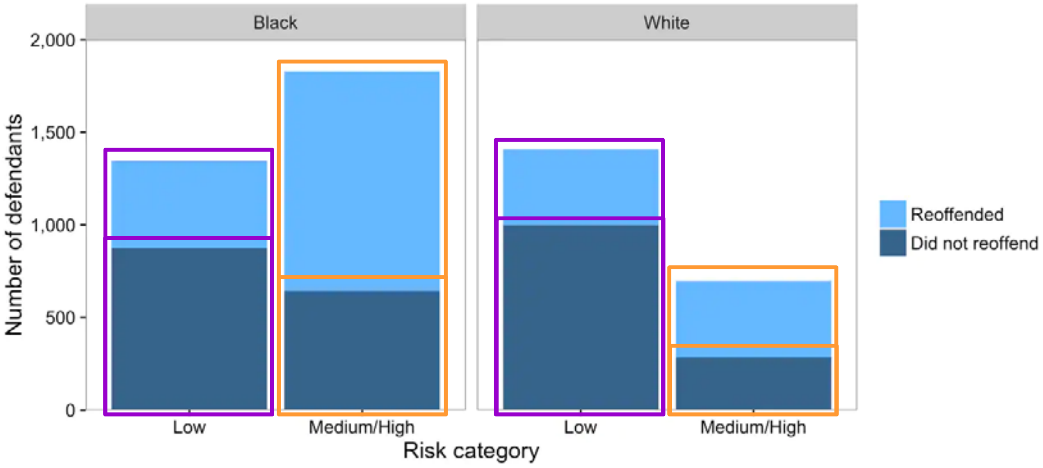


Figure 3. Evaluation of Predictive Parity in the COMPAS Algorithm.

Figure 3 illustrates the second criterion, *predictive parity*. Look at the Low columns highlighted in purple in the left and right parts of the figure. These columns correspond to individuals for whom the predicted risk $\hat{Y}$ is low. To achieve prediction parity, the proportion of the two areas in different colors representing the actual outcome Y should be the same in both columns. The same must apply to the Medium/High columns, which are highlighted in orange. If this condition is met, the prediction is independent of the sensitive feature *A*. With some indulgence, one could argue that Figure 3 shows that the criterion of predictive parity is fulfilled, or at least that the imbalances are much less pronounced compared to the previous case.

FIGURE 4

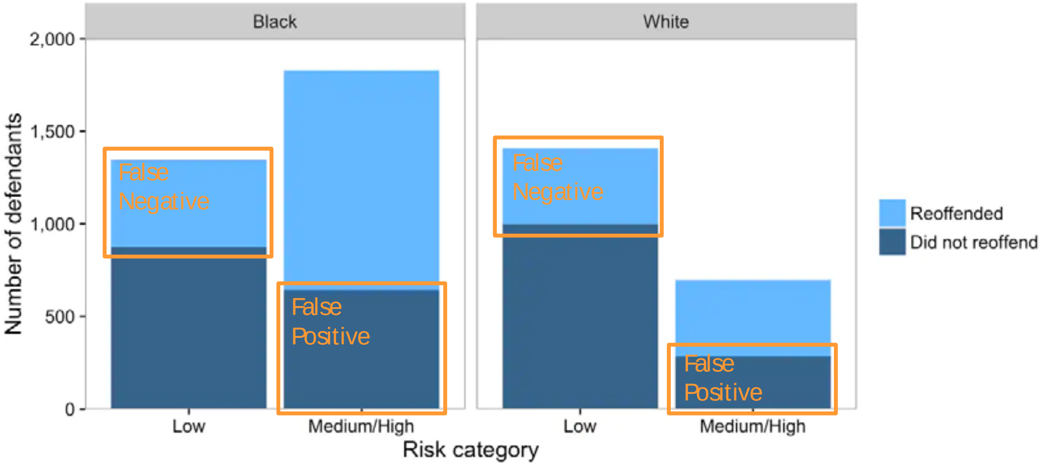


Figure 4. Evaluation of Positive and Negative Error Parity in the COMPAS Algorithm.

Figure 4 illustrates the third criterion, namely the *parity of positive and negative errors*. For this criterion to be met, the ratio of False-Positive to False-Negative errors on the left and right sides should be the same. The column segments to be compared, which correspond to the prediction errors, are highlighted in orange. The figure shows that COMPAS makes proportionally more false-positive errors on the left-hand side, i.e. predicts a medium to high risk of people not reoffending, and more false-negative errors on the right-hand side.

This means that the COMPAS algorithm does not fulfil two of the three criteria. However, is it possible for an algorithm to fulfil all three criteria simultaneously? A theoretical result in mathematical statistics states that an algorithm based on observations, i.e. real data, can only fulfil the three criteria simultaneously *if the real data has no bias*. Specifically, the following condition must be fulfilled in the actual data: The probability of the outcome Y = 1 (reoffending) must be the same for all values of the sensitive attribute A, i.e.:

$$P(Y=1 \mid A=a) = P(Y=1 \mid A=b)$$

for any values a and b of the sensitive variable A. In view of the above, the COMPAS algorithm could be defended with the argument that distortions are present in the real data and that the algorithm therefore fulfils at least the second criterion under the given circumstances.

11. The complexity of detecting data bias

Let us now look at another case study involving a publicly available dataset that is often used for testing purposes. This dataset relates to undergraduate (equivalent to a bachelor's degree) admissions outcomes at the University of California, Berkeley, in 1973, and includes anonymized data on the major choice and admissions outcomes of 12,763 applicants who applied that year. It also records the gender of the applicants, categorized as male or female. It is important to note that no algorithms were used at the time. The results recorded in the data set were exclusively the result of human decisions. As an aside, the Berkeley Admissions Dataset is a well-known example of the Simpson paradox, a phenomenon in which aggregate data can obscure or reverse trends within subgroups, leading to potentially misleading interpretations (Bickel *et al.* 1975).

FIGURE 5

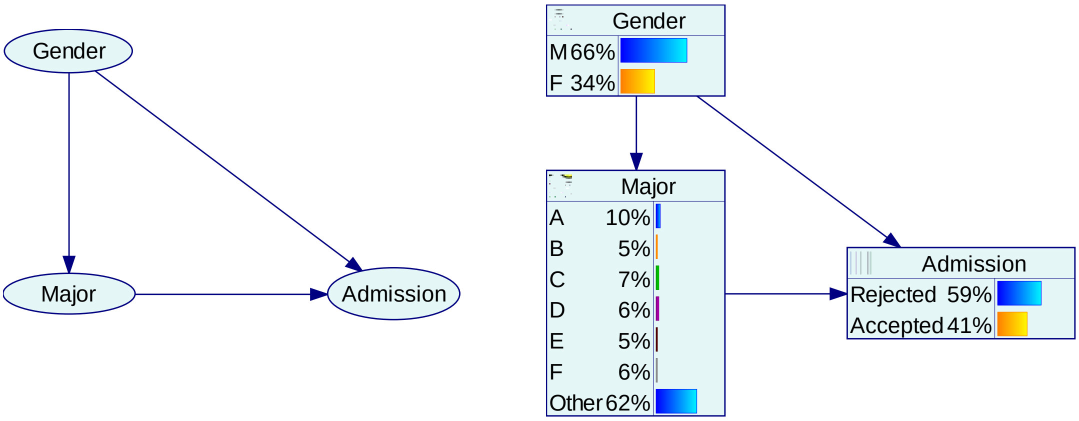


Figure 5. Structure and Statistical Distribution of Data, Undergraduate Admissions (Major) at UC Berkeley, 1973.

On the right, Figure 5 shows the statistical frequencies observed in the dataset. To ensure greater anonymity, the names of the choices have been replaced by letters (A, B, C, D, etc.). The calculations show that the admission rates differ between the genders: 44% for men and 35% for women. At first glance, one might conclude that the dataset shows a clear bias in favour of men in admissions decisions. However, the situation is more complex. Further statistical analysis shows that: a) admission rates vary across majors; b) women, who are likely to be better prepared, tend to choose more challenging faculties that have lower admission rates.

Let us create a probabilistic predictor using the Berkeley dataset. A common approach is to first define a network structure as shown on the left side of Figure 5, where the arrows represent causal influences between the variables involved (Pearl *et al.* 2016). Given the analyses conducted, the presence of some arrows is obvious: gender can influence the choice of major, and the choice of major can influence the likelihood of admission. However, the arrow from gender to admission is more problematic: ideally, there should be no influence at all, i.e. the evaluator (whether human or algorithm) should be impartial regarding gender.

If we omit some technical details, we can construct a probabilistic predictor that uses the relative frequencies in the dataset as probability values. By combining the network structure with these derived probability values, we can basically create a probabilistic predictor. Specify your gender (characteristic A) and your chosen major X, and the predictor estimates your probability of being admitted $\tilde{Y}$. The actual result, once known, will be Y. On closer inspection, however, it becomes clear that this predictor does not fulfil any of the previously introduced fairness criteria regarding gender A. It would therefore be rather unattractive as an algorithm. Modifications could be considered to improve its impartiality, but how? Adjusting the numerical probability values to achieve fair predictions is of course possible, but if the predictor deviates from the actual data, even if it contains bias, what rational justification can it have?

FIGURE 6

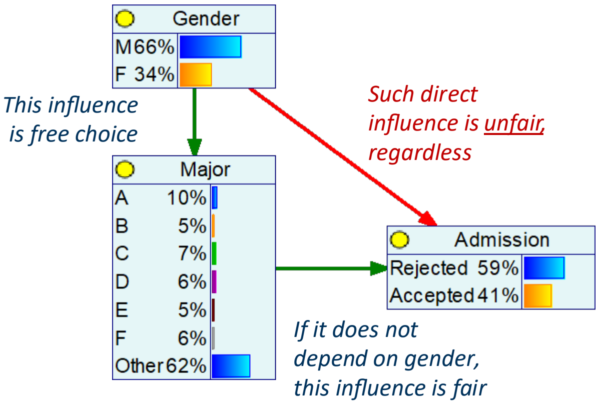


Figure 6. Bias depends on the causal path considered.

12. Enhancing algorithmic fairness through counterfactual analysis

The recent academic literature offers many suggestions on how to tackle this problem. One particularly interesting approach from an intuitive point of view is described by Chiappa (2019). Figure 6 summarizes the earlier discussion. The pathway from gender to admission outcome via choice of major should not be considered biased. In contrast, the direct pathway from gender to admission outcome is considered unfair and ethically unacceptable. The first step in addressing this issue is to assess the relative impact of each pathway on the imbalance in admission rates.

The proposed method for identifying bias is based on counterfactual reasoning. Intuitively, it is as if a female applicant were to ask herself, "What if I had shown up for the admission interview as a man?" Such experiments are indeed carried out in the field of social psychology. Apart from the practical difficulties, however, a real experiment would only produce one result, whereas two are required. What would have happened if the applicant had introduced herself as a woman? And what if she had presented as a man? In an ethically desirable scenario, the result should be the same in both cases. In practice, however, it is impossible to conduct the exact same interview with two different appearances.

FIGURE 7

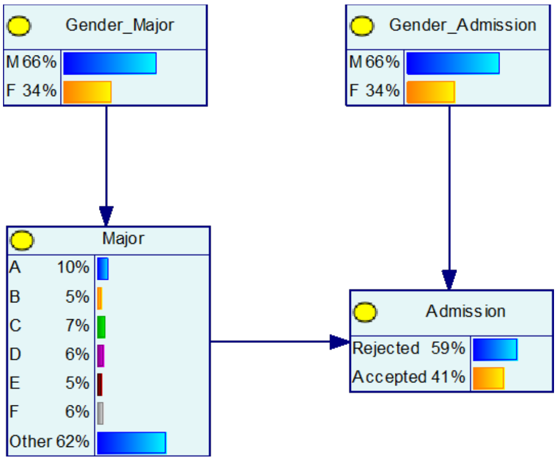


Figure 7. Probabilistic structure (causal network) for estimating the counterfactual.

Figure 7 illustrates a technique that, when certain technical criteria are met, allows the use of a probabilistic predictor derived from data to determine the relative probabilities in both scenarios, effectively creating a ‘double virtual experiment’. In this figure, the gender-specific data have been split for each of the two paths. For example, a person may can be treated as a woman when choosing a field of study (the first pathway) and as a man when deciding on admission (the second pathway).

FIGURE 8

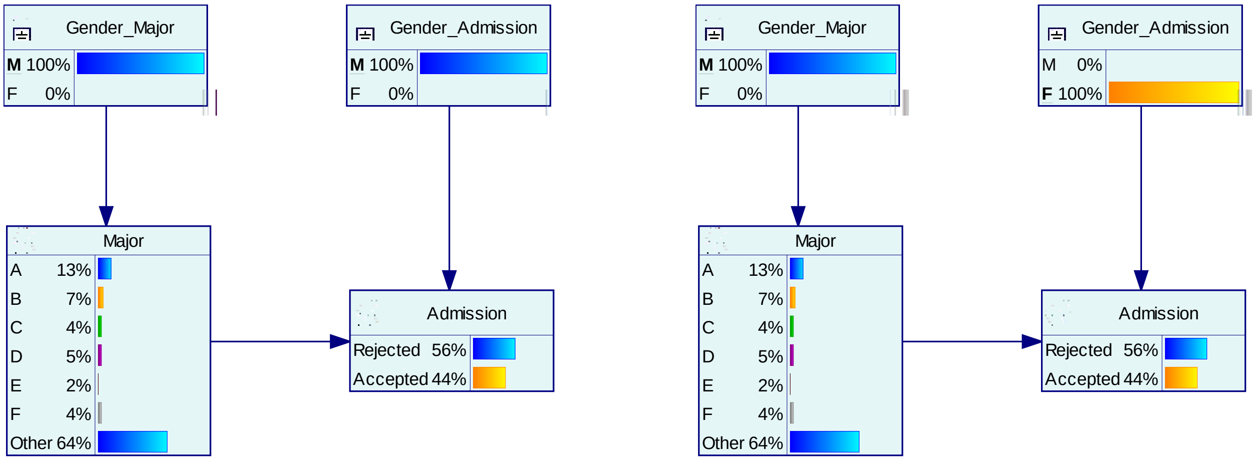


Figure 8. Estimating the counterfactual for male candidates.

The two parts of Figure 8 show the results of the algorithm when the subject is male for the first path and either male (left) or female (right) for the second path. As you can see, the predicted admission rate in both parts of Figure 8 is identical at 44%. Therefore, there is no bias in this scenario.

FIGURE 9

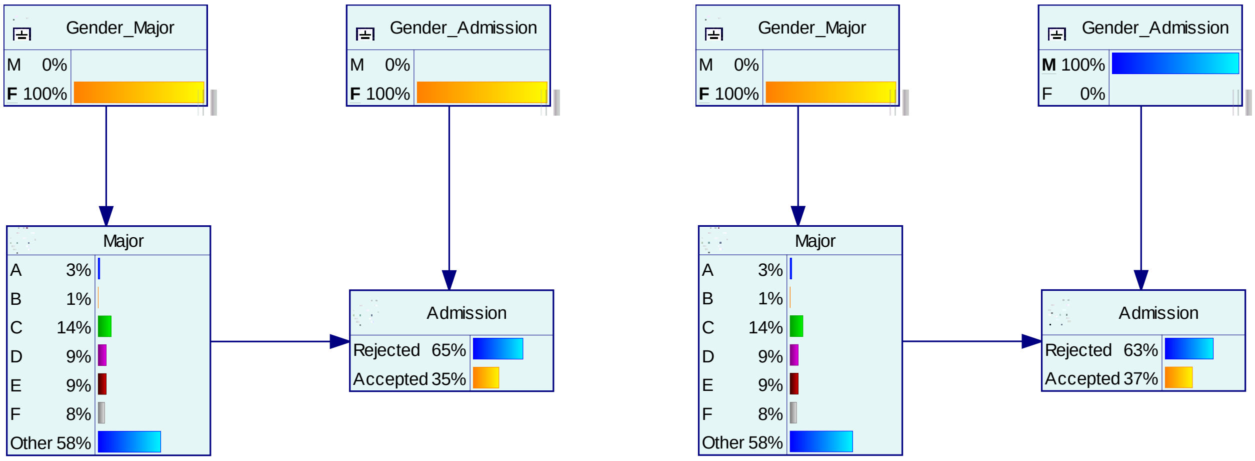


Figure 9. Estimating the counterfactual for male candidates.

The two parts of Figure 9 show the results of the predictor for the opposite counterfactual situation: when the subject is female for the first path and either female (left) or male (right) for the second path. In this second counterfactual situation, the admission rates are different and are 35% and 37% respectively.

To summarize, we can infer from this analysis that: a) there is indeed a bias in the dataset that disadvantages women; b) the bias is much less pronounced than it initially appeared when comparing the absolute admission rates of 44% for men and 35% for women. In other words, this analysis shows that a significant portion of the difference is due to the tendency of women to apply for subjects with lower admission rates. Nonetheless, there remains a component of bias that is unfair, and this component should be corrected in the predictor to obtain an unbiased algorithm.

Given this reasoning, the technical challenge in debiasing a predictor is to stick to probability values that are consistent with the actual observed data while correcting for those that introduce bias, as determined by the counterfactual analysis. In other words, the goal is to make a 'surgical' correction, that affects only the 'pathological' aspect of the algorithm without completely jeopardizing its consistency with the real data of the dataset. In general, such a correction is a technical-mathematical problem that is not always easy to solve, since in more complex cases the fair and unfair paths may partially overlap. The positive aspect is that the search for a solution can be shifted to the technical area once the method has been agreed.

Finally, it is important to emphasise that the case studies presented here focus on biases in the source data that ultimately affect the algorithm's predictions. This problem becomes even more critical when advanced prediction models are used (such as those of deep learning; Bishop and Bishop 2023), where additional layers of bias may occur. Due to space limitations, we have not analysed these potential sources of bias or discussed possible mitigation strategies, such as those offered by Explainable AI (XAI).

13. Towards better bias detection and debiasing in algorithms

The case studies discussed in this article and many other cases that have received a lot of media attention illustrate two important points. On the one hand, they demonstrate the instrumental value of technology and expertise by highlighting the complexity of formal and quantitative bias detection. On the other hand, they highlight the need for further maturation and development in this area. From this perspective, it is commendable that the recent proposal for a European regulation on harmonized rules for artificial intelligence takes a risk-based approach and classifies certain categories of applications as high-risk and in need of special treatment. The requirements for this treatment include ex-ante validation prior to deployment, continuous monitoring, and human supervision. The intention of the legislator can be seen as a positive push towards further technological advances that could promote the impartiality and fairness of algorithms. At the same time, it is important to bear in mind that this is only possible if the procedures used also make it possible to recognise and remove cultural, political or economic barriers that could affect the quality of the whole process.

Further reflection on the relationship between bias and debiasing techniques in the field of human and artificial intelligence is certainly important for progress in the quality of reasoning and decision-making. A prerequisite for this is always that we know biases and debiasing techniques in the human and artificial domains often do not work in a completely analogue way and therefore cannot be directly transferred from one domain to the other. Nevertheless, the examples discussed in this document show that to fully implement the guidelines set out in the EU AI Act, it is essential that we improve our understanding of bias and develop accurate debiasing techniques. This can ensure that automated reasoning processes achieve a level of fairness and neutrality that is compatible with ethical and social acceptability.

14. Conclusion

When we reason or make decisions, various cognitive, emotional, and motivational factors can lead us to resort to mental shortcuts that allow us to move forward more quickly, but which can also lead to biases, i.e. systematic distortions in our reasoning process. In these cases, we can resort to various debiasing strategies, the effectiveness of which can be tested through experimental research. In the field of justice, it is also appropriate to work towards ensuring that the decision-making procedures codified by the rules, case law and custom are coherent and do not contradict the possibilities, but also the limits, of our cognitive system.

Artificial intelligence algorithms can also exhibit biases. However, the investigation of these occurrences shows that such biases is not due to an inherent 'unpredictable subjectivity' of the algorithms. Instead, it becomes clear that a thorough analysis of the formal methods is necessary to achieve a satisfactory level of methodological awareness. It is possible to disarm the algorithms with appropriate modifications and human-programmed control systems. This means intervening to ensure that the predictions and actions generated by artificial intelligence are consistent with reality and in line with the goals for which the algorithms were developed.

In both human and artificial intelligence, debiasing strategies reduce the likelihood of bias and thus increase the quality of reasoning and decisions made.

References

Alicke, M.D., *et al.*, 2008. Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin* [online], 34(10), 1371–1381. Available at: https://doi.org/10.1177/0146167208321594

Angwin, J., and Larson, J., 2015. Machine Bias. *ProPublica* [online], 23 May. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bickel, P.J., Hammel, E.A., and O'Connell, J.W., 1975. Sex Bias in Graduate Admissions: Data from Berkeley. *Science* [online], 187(4175), 398-404. Available at: https://doi.org/10.1126/science.187.4175.398

Bishop, C. M., and Bishop, H., 2023. *Deep Learning: Foundations and Concepts* [online]. Cham: Springer Nature. Available at: https://doi.org/10.1007/978-3-031-45468-4

Catellani, P., 2010. Fatti e controfatti nel ragionamento giudiziario. *Sistemi Intelligenti*, 22(3), 209–222. Available at: https://www.rivisteweb.it/doi/10.1422/32620

Chiappa, S., 2019. Path-Specific Counterfactual Fairness. *Proceedings of the AAAI Conference on Artificial Intelligence* [online], 33(01). Available at: https://doi.org/10.1609/aaai.v33i01.33017801

Corbett-Davies, S., *et al.*, 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*, 17 October.

Dale, S., 2015. Heuristics and biases: The science of decision-making. *Business Information Review* [online], 32(2), 93–99. Available at: https://doi.org/10.1177/0266382115592536

Damasio, A.R., 1996. The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* [online], 351(1846), 1413–1420. Available at: https://doi.org/10.1098/rstb.1996.0125

Danziger, S., Levav, J., and Avnaim–Pesso, L., 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* [online], 108(17), 6889–6892. Available at: https://doi.org/10.1073/pnas.1018033108

Dressel, J., and Farid, H., 2016. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* [online], (4)1, January. Available at: https://doi.org/10.1126/sciadv.aao5580

Gawronski, B., *et al*., 2018. Effects of incidental emotions on moral dilemma judgments: An analysis using the CNI model. *Emotion* [online], 18(7), 989–1008. Available at: https://doi.org/10.1037/emo0000399

Gigerenzer, G., 2008. *Gut feelings. Short cuts to better decision making*. London: Penguin.

Hertwig, R., and Grüne–Yanoff, T., 2017. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science* [online], 12(5), 973–986. Available at: https://doi.org/10.1177/1745691617702496

Inzlicht, M., Legault, L., and Teper, R., 2014. Exploring the mechanisms of self–control improvement. *Current Directions in Psychological Science* [online], 23(4), 302–307. Available at: https://doi.org/10.1177/0963721414534256

Kahneman, D., 2011. *Thinking, fast and slow*. New York: FSG Adult.

Kirkpatrick, K., 2017. It's not the algorithm, it's the data. *Communications of the ACM* [online], 60(2). Available at: https://doi.org/10.1145/3022181

MacDougall, C., and Baum, F., 1997. The devil's advocate: A strategy to avoid groupthink and stimulate discussion in focus groups. *Qualitative Health Research* [online], 7(4), 532–541. Available at: https://doi.org/10.1177/104973239700700407

Martin, J.W., and Cushman, F., 2016. Why we forgive what can't be controlled. *Cognition* [online], 147(4), 133–148. Available at: https://doi.org/10.1016/j.cognition.2015.11.008

McClure, J., Hilton, D.J., and Sutton, R.M., 2007. Judgments of voluntary and physical causes in causal chains: Probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology* [online], 37(6), 879–901. Available at: https://doi.org/10.1002/ejsp.394

Metcalfe, J., and Shimamura, A.P., 1994. *Metacognition: Knowing about knowing* [online]. Cambridge, MA: MIT Press. Available at: https://doi.org/10.7551/mitpress/4561.001.0001

Pearl, J., Glymour, M., and Jewell, N.P., 2016. *Causal Inference in Statistics: A Primer*. Hoboken: Wiley.

Roese, N.J., and Olson, J.M., 1996. Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology* [online], 32(3), 197–227. Available at: https://doi.org/10.1006/jesp.1996.0010

Savioni, L., and Triberti, S., 2020. Cognitive biases in chronic illness and their impact on patients' commitment. *Frontiers in Psychology* [online], 11, 579455. Available at: https://doi.org/10.3389/fpsyg.2020.579455

Smith, C., and Greene, E., 2005. Conduct and its consequences: Attempts at debiasing jury judgments. *Law and Human Behavior* [online], 29(5), 505–526. Available at: https://doi.org/10.1007/s10979-005-5692-5

Smith, M., 2016. In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures. *The New York Times*, 22 June.