CONTROLLING BIAS IN MACHINE LEARNING: MITIGATING HUMAN INFLUENCES ON ALGORITHMIC DECISION-MAKING

MARCEL PITERMAN

Universidade Católica Portuguesa, Faculty of Law,
Católica Research Centre for the Future of Law, Portugal

ISSN 1392-0588 (spausdintas)
ISSN 2335-8769 (internetinis)
https://doi.org/10.7220/2335-8769.79.5
2023. 79

SUMMARY. In this essay, it is assumed that every human being's activity can be influenced by external circumstances that should not impact decision-making, the so-called biases. Cognitive biases were systematized in order to identify heuristic processes with the unconscious objective of reducing the complexity of tasks, which fatally lead to systematic logical errors. In addition, humans tend to obey an authoritative figure, even if the authority instructs them to perform acts conflicting with their personal conscience, as was found in the Milgram experiment where a very high proportion of people would fully obey the instructions given. So, when machine learning involves information provided by humans to algorithms, considering that this information may have been biased or subjected to personally conflicting instructions, ways of controlling the algorithmic results and the data initially provided by humans must be developed.

KEYWORDS: machine learning, bias, algorithms, Milgram experiment, decision-making.

EXTERNAL FACTORS THAT SHOULD NOT INFLUENCE DECISION-MAKING

Judicial decisions require rigorous reasoning and explanation, but research shows that irrelevant and unconscious factors can distort the decision-making process. This has raised concerns about the potential impact of cognitive biases on human decisions, including judicial decisions.[1] So, it must now be seen to be proven that, in many circumstances, judges can be as susceptible as lay people to systemic weaknesses in their cognitive decision-making apparatus.[2]

If this is so, then other studies regarding biases in human decision-making, while not conducted in courtrooms or close simulations with judges as subjects, should give us further concern about how reasoning and decision-making might be influenced at the sub-rational level by circumstance or by skilled manipulation.

[1] See Craig E. Jones. The Troubling New Science of Legal Persuasion: Heuristics and Biases in Judicial Decision-Making, *Advocates' Quarterly*, v. 41, n. 1, 2013. See also Shai Danziger, Jonathan Levav, Liora Avnaim-Pesso and Daniel Kahneman. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, v. 108, n. 17, 2011.

[2] See Andrew J. Wistrich, Chris Guthrie, Jeffrey J. Rachlinski. Can judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding, *Pennsylvania Law Review*, v. 153, 2005.

So, other biases might affect the accuracy of decisions, and decision-making can be sub-rationally influenced by all manner of supposedly irrelevant inputs, such as political orientation.

These biases appear to arise due to what psychologists refer to as heuristics – cognitive shortcuts that are used as defaults in the decision-making process.[3] These heuristics operate mostly at a sub-conscious level, only occasionally connected with the simultaneous, rational thought process going on above the cognitive waterline (Jones 2013: 50).

Digital marketing experts have become adept at using heuristics, or mental shortcuts, to influence consumer behavior through targeted advertising. However, recent research has shown that these heuristics can also lead to irrational decision-making, and that this effect is not limited to laypeople but also affects judges and other experts. This is particularly concerning because these heuristics can be systematically exploited by algorithms, which can influence our behavior in ways that are invisible and difficult to resist (Jones 2013: 54).

For example, the algorithms used by social media platforms and search engines are designed to optimize engagement and increase time spent on the platform. To achieve this, they may present users with content that reinforces their pre-existing beliefs or interests, creating "echo chambers" that can polarize public opinion and exacerbate social divisions. Additionally, these algorithms may use data such as location, search history, and demographic information to tailor advertising and content to individual users, potentially leading to discriminatory or exploitative outcomes.

Therefore, it is important to be aware of the potential impact of algorithmic inputs and outputs on social behavior and to work towards developing more ethical and transparent algorithms that consider the potential for heuristics to bias decision-making.

As researchers continue to gain a better understanding of heuristics and their manipulation, it is crucial to consider the actions individuals can take in response to these phenomena. As a result, there is a growing urgency to discuss these and other related questions as this new field of science develops.

HEURISTICS AND BIASES: GENERAL CONCEPTS AND CHARACTERISTICS

It was through the work of Daniel Kahneman and Amos Tversky that the subject was introduced in the field of psychology. The authors proposed that judgments

[3] The name of the famous computer in Arthur C. Clarke's 1968 novel *2001: A Space Odyssey*, HAL-9000, was an amalgam of "Heuristic/Algorithmic" because that computer, like the human mind, used both processes.

made in conditions of uncertainty were often the result of simple cognitive processes – heuristics – that worked reasonably well most of the time, particularly in the simpler, less information-rich society in which our brains evolved. When they were applied to the many difficult situations of modern existence, however, they tended to produce a pattern of systematic errors, or biases (Tversky Kahneman 1974: 1124).

Social psychologists began explaining phenomena such as stock market bubbles and wars by observing how people responded to "trigger" stimuli, rather than assuming that they always acted in their rational self-interest or based on other measures of utility (Jones 2013: 56).

Modern analyses divide decision-making processes into two systems: intuitive and deliberative. Kahneman called the former "System 1" and the latter "System 2" and suggested that the former, characterized by high speed, high automaticity, low effort, low awareness, and low conscious control, might or might not be overridden by the latter, "systematic" processes which have the opposite characteristics. System 1 and System 2 are two distinct cognitive processes. These two systems are responsible for the way people think and make decisions. System 1 is a fast, automatic, intuitive, and effortless mode of thinking. It is responsible for quick and instinctive reactions to the world around us, and is driven by past experiences and automatic associations. For example, recognizing a familiar face or reacting quickly to a sudden loud noise are functions of System 1. System 2, on the other hand, is a slower, more deliberate, analytical, and effortful mode of thinking. It involves conscious mental effort, attention, and reasoning to solve problems and make decisions. For example, solving a difficult math problem or analyzing a complex piece of information requires System 2 thinking (Kahneman 2011: 17).

While both System 1 and System 2 are important and necessary for people's daily lives, they have different strengths and weaknesses. System 1 thinking is efficient and fast, but it can also be prone to biases and errors, particularly when people rely too heavily on their past experiences and automatic associations. System 2 thinking, on the other hand, is more accurate and reliable, but it requires more effort and attention.

One important aspect of understanding System 1 and System 2 is recognizing when people are using each mode of thinking and using them appropriately. For example, in situations that require quick reactions and immediate decisions, such as driving in traffic or playing sports, System 1 thinking is more appropriate. In contrast, in situations that require careful analysis and problem-solving, such as planning a project or making a financial decision, System 2 thinking is more appropriate.

It is even easier to understand System 1 and System 2 through an example that Robert Cialdini describes as a "reciprocity" bias. Humans, intensely social animals, seem to be programmed with a heuristic that triggers reciprocity or compliance when a favor, even a small one, is done. In one study, researchers found that waiters who gave diners a mint after their meal increased their tips by 3%. However, when the waiters gave diners two mints, and then walked away before the diners could take them, tips increased by 14%. This effect was explained by the "reciprocity" bias: when people receive something, they feel obligated to give something in return, even if it is just a small favor like leaving a larger tip (Cialdini 1993: 33–36). This is a clear example of how the reciprocity bias can be used to influence people's behavior.

This simple reciprocity technique is exploited by telemarketers, charities, and businesses as most people have experienced some similar episode of conflict between System 1 and System 2 minds.

So, when someone makes automatic decisions based on the primitive System 1, they often do not recognize them as such. They rarely admit, even to themselves, that there was anything but a perfectly reasonable explanation for what they have done. Once the subconscious decision has been made, and assuming that it is not "caught" by a rational re-think, the conscious mind constructs its explanations, which range from quite simple to remarkably elaborate. These explanations may have little to do with the real reason behind the decision, but to the decision-maker, they are the gospel truth. And even when an "automatic," System 1 decision is subject to a rational review as evidence is gathered, the mind's inclination is to support and confirm, rather than to critically analyze and constantly reconsider (Jones 2013: 60–61).

FRAMING AND REPRESENTATIVENESS ERRORS

Framing is a method of changing analysis by structuring the question in a different way (Kahneman Tversky 1984: 3). For instance, psychologists and economists have long observed that people consider amounts "framed" as losses instead of gains to be more significant: make the point that loss aversion underlies their beliefs about fairness: most people think that a store is behaving "unfairly" if it increases the price of a snow shovel after a blizzard, but not if it reduces the price after a stretch of warm weather. The difference, from a fairness point of view, appears to be at what point in time the price is "framed."

The representativeness heuristic was first introduced by Amos Tversky and Daniel Kahneman in their seminal 1974 paper "Judgement under Uncertainty: Heuristics and Biases." They proposed that people often rely on stereotypes and prototypes to make judgments about events or people, leading to errors in judgement (Tversky Kahneman 1974: 1124). The concept of representativeness has since been studied extensively in the fields of cognitive psychology and behavioral economics,[4] and is now widely recognized as a common heuristic process that people use to make judgments.

The representativeness heuristic is a mental shortcut that people use to make judgments about the probability of an event based on how similar it is to typical examples. This heuristic assumes that the more representative an event or person is of a category, the more likely it is to belong to that category. For example, if a person sees someone wearing a lab coat and carrying a clipboard, they may assume that the person is a scientist, even if they have no other evidence to support that assumption (Tversky Kahneman 1974: 3).

In other words, the representativeness heuristic is a shortcut whereby people form a view based on a stereotype rather than a true probabilistic assessment. It can be strongly influenced by the introduction of meaningless evidence. This bias can be helpful in some cases, but it can also lead to stereotyping and prejudice.

AVAILABILITY

The availability heuristic was also introduced by Tversky and Kahneman in their 1974 paper. They proposed that people tend to judge the probability of an event based on how easily they can recall similar events from memory. This heuristic has since been studied extensively in the fields of cognitive psychology, social psychology, and neuroscience, and has been shown to have a significant impact on decision-making in a variety of domains, including politics, health, and finance (Tversky Kahneman 1974: 15).

The availability heuristic is a mental shortcut that people use to make judgments about the probability of an event based on how easily it comes to mind. This heuristic assumes that the more available or easily retrievable an example is from memory, the more likely it is to occur. For example, if a person is asked to name a type of fruit, they may be more likely to say “apple” if it is the first fruit that comes to mind, even if other fruits are more common.

[4] See Clara Martins Pereira. Reviewing the literature on behavioral economics. *Capital Markets Law Journal*, v. 11, n. 3, 2016, 414–428. The article provides an overview of some of the most significant literature on behavioral economics and its importance in financial decision-making. In particular, this literature review looks at the needs of particularly vulnerable consumers and at the different regulatory strategies adopted to ensure their protection.

ADJUSTMENT AND ANCHORING

Tversky and Kahneman expanded upon this concept in their 1974 paper, showing that people tend to start from an initial value (the anchor) and adjust it based on new information. This heuristic has since been studied extensively in the fields of cognitive psychology and behavioral economics and has been shown to have a significant impact on decision-making in a variety of domains including negotiation, pricing, and judgment of fairness (Tversky Kahneman 1974: 20).

It is a powerful influence that manifests when people are dealing with numbers. Put simply, numerical anchoring means that if people are asked to come up with a value, they will be influenced by numbers they have recently seen. Thus, they become "anchored" to a number and it skews subsequent estimates towards it, even if the anchoring number has (or at least, should have) no relationship to the value being calculated. The evolutionary origin of anchoring may be related to the fact that, in a primitive environment, before the development of written numbers, holding numerical values in people's heads and focusing on small changes to them was the most important mathematical task faced, and, because they were not likely to confront a series of entirely discrete calculations in rapid succession, such anchoring had few downsides. In other words, like all heuristics that survived the evolutionary process, it has worked well enough, enough of the time.

In summary, the adjustment and anchoring bias is a mental shortcut that people use to make judgments by starting from an initial value (the anchor) and then adjusting that value based on new information. This heuristic assumes that people tend to be biased toward the initial anchor, even if it is arbitrary or irrelevant. For example, if a person is asked to estimate the price of a product, and the initial price they are given is high, they may tend to estimate a higher price than if they were given a lower initial price.

CONFIRMATION BIAS

Confirmation bias is the tendency to search for, interpret, and remember information in a way that confirms people's pre-existing beliefs or hypotheses, while disregarding information that contradicts them. This bias was first identified in the field of psychology by Peter Wason in the 1960s. Since then, many studies have confirmed the existence of confirmation bias, which has been shown to have a significant impact on decision-making in a variety of domains, including politics, economics, science, and law.

For example, in a study on jury decision-making, researchers found that jurors who had strong preconceptions about a defendant's guilt or innocence tended to selectively focus on evidence that supported their pre-existing beliefs, while disregarding evidence that contradicted them (Rassin Eerland Kuijpers 2010: 231–246). This bias can lead to incorrect judgments and erroneous decisions, and it is important to be aware of it and actively seek out and consider alternative perspectives and evidence.

HINDSIGHT BIAS

The human mind is not optimally designed to comprehend the complexities of the world. Instead, it evolved to quickly navigate and survive challenges while ensuring the continuation of the species. If humans were meant to fully understand everything, their minds would need a machine to accurately replay past events, but it would slow down so much that it would become difficult to operate effectively. Psychologists refer to the phenomenon of overestimating one's knowledge at the time of an event due to subsequent information as the "hindsight bias," often manifesting as the "I knew it all along" effect. The civil servant in question deemed the trades that resulted in losses as "gross mistakes," a term frequently used by journalists to describe decisions that cost a candidate an election. However, labeling such decisions as mistakes should be based on the information available at the time of the decision, not on subsequent knowledge. A more dangerous consequence of this bias is that individuals who excel at hindsight may erroneously believe they can predict the future, leading to unwarranted confidence in their predictive abilities. As a result, even catastrophic events such as the 9/11 attacks fail to convince us that we live in a world where significant events are inherently unpredictable (Taleb 2004: 56).

Hindsight bias is the tendency to believe, after an event has occurred, that we would have predicted or expected it to happen. This bias was first identified in the field of psychology by Baruch Fischhoff in the 1970s. He explains that hindsight bias occurs because people tend to think about what they already know and then work backwards to what they did not know at the time. They do this unconsciously, without realizing that they are doing it. As a result, they overestimate their ability to have predicted what happened based on the information they had at the time. This can lead to a false sense of confidence in their ability to predict future events (Fischoff 1975: 288–299).[5]

[5] According to the author, "When we attempt to understand past events, we implicitly test the hypotheses or rules we use to both interpret and anticipate the world around us. If, in hindsight, we systematically underestimate the surprises which the past held and holds for us, we are subjecting those hypotheses to inordinately weak tests and, presumably, finding little reason to change them. Thus, the very outcome knowledge which gives us the feeling that we understand what the past was all about may prevent us from learning anything from it."

Since then, many studies have confirmed the existence of hindsight bias, which can lead to overconfidence in our own judgments and decisions. For example, in a study on medical decision-making, researchers found that doctors who were given information about a patient's condition after the diagnosis tended to believe that they would have made the correct diagnosis even without the additional information. This bias can lead to errors in judgement and decision-making, and it is important to be aware of it and actively seek out information that can help us make better decisions.

Hindsight bias has been shown to occur in the courtroom as well, mainly in liability cases. In such cases, the task of the judges or jurors is to assess how foreseeable an outcome was and to evaluate whether the plaintiff's behavior took this risk into consideration. The problem is that judges evaluate the outcome in hindsight, while the plaintiff only had the chance to provide foresight about it. For example, in one case a physician was accused of malpractice because he failed to detect a tiny tumor in an early chest radiography. The tumor got bigger and the patient died as a result, leading to the malpractice claim. The physician was found guilty after another radiologist, who saw the radiographs after the tumor was found, testified that the tumor could have been detected in the early radiography. Clearly, the second radiologist had the benefit of knowing the tumor was actually there, an advantage the first physician did not have at the time. In another example, judges who were informed that a psychiatric patient became violent were more likely to find the patient's therapist negligent than those who did not receive information about the outcome and its severity (Peer Gamliel 2013: 115).

CONJUNCTION FALLACY

Conjunction fallacy is the tendency to believe that the co-occurrence of two events is more likely than the occurrence of either event alone. This bias was first identified in the field of psychology by Amos Tversky and Daniel Kahneman in the eighties. The authors showed that when subjects are asked to rate the likelihood of several alternatives, including single and joint events, they often make a "conjunction fallacy." That is, they rate the conjunction of two events as being more likely than one of the constituent events (Tversky Kahneman 1983: 293–315). Since then, many studies have confirmed the existence of the conjunction fallacy, which can lead to erroneous judgments and decisions.

For example, in a study on probability judgement, researchers found that participants were more likely to believe that a description of a person was accurate if it included more details, even if those details made the description less likely to be true. This bias can lead to an overestimation of the probability of complex events and is important to be aware of when making decisions based on probabilities.

This type of judgmental bias also relates to how people judge the probability of events based on the detail in which these events are described. It has been found that more detailed descriptions of an event can give rise to higher judged probabilities. This bias has been called the conjunction fallacy because it shows that people erroneously believe that events described in more detail are more probable than those that are described in less detail. According to classic probability theory, less detailed events contain various instances of more detailed events and thus cannot be less probable than any of the contained events. For example, just as the probability of an object being a fruit cannot be smaller than the probability of a suspect being convicted of a crime cannot be smaller than the probability that he will be convicted of a specific crime, such as burglary (Peer Gamliel 2013: 116).

BIASED DECISIONS IN SEQUENTIAL RULING

Biased decisions in sequential ruling occur when the order in which decisions are made affects the overall outcome. This bias was first identified in the field of law by Cass Sunstein in the book "One Case at a Time – Judicial Minimalism on the Supreme Court." He identified a bias in sequential ruling that he called the "availability cascade." This bias occurs when a judge or decision-maker makes a ruling based on the previous ruling without fully considering the underlying facts and legal principles. Essentially, the decision-maker is persuaded by the availability and influence of the previous ruling, rather than independently considering the case at hand. For example, imagine a judge hears a case where the plaintiff sues for damages due to a car accident. The judge rules in favor of the plaintiff and awards damages. Later, the same judge hears a similar case with similar facts, but, this time, the defendant argues that the previous ruling should be overturned or discounted. However, the judge may be influenced by the previous ruling and feel compelled to make a similar ruling, even if the facts and legal principles suggest otherwise.

Since then, many studies have confirmed the existence of this bias, which can lead to unfair and inconsistent decisions. For example, in a study on parole decisions, researchers found that judges were more likely to grant parole early in the day or after a meal break, and less likely to grant parole later in the day or before a meal break. This bias can lead to inconsistent and unfair decisions and is important to be aware of when making sequential decisions. Also, when judges make repeated sequential rulings, they tend to rule more in favor of the *status quo* over time, but they can overcome this tendency by taking a food break (Danziger Levav Avnaim-Pesso Kahneman 2011: 6892).

BIASES AND THE MILGRAM EXPERIMENT

The Milgram experiment is a classic example of how cognitive biases can impact decision-making and how they can lead to unethical behavior. The experiment was a series of social psychology experiments conducted by Stanley Milgram in the sixties. The experiments aimed to investigate the willingness of participants to obey an authority figure, even when their actions conflicted with their personal beliefs and values. Participants were instructed to administer electric shocks to another person, who was a confederate in the experiment. The shocks were not real, but the participants did not know this. The shocks were meant to increase in intensity with each incorrect answer given by the confederate. Despite the confederate's pleas and screams of pain, many participants continued to administer shocks as they were instructed by the experimenter, who was an authority figure in a lab coat.

The experiment demonstrated the power of situational factors and the influence of authority figures on human behavior. It also raised ethical concerns regarding the use of deception and psychological harm to participants in research.

What will be discussed next is how some of the common cognitive biases discussed earlier can be related to the Milgram experiment. The representativeness heuristic can influence how the participants in the Milgram experiment were selected. If the participants were not representative of the population or the problem being studied, the results may not be generalizable to other contexts. For example, if the participants were predominantly from a certain demographic, such as college students, the results may not be applicable to other age groups or populations. The availability heuristic can impact how the participants in the Milgram experiment were chosen. If certain individuals were more easily available or accessible, they may have been overrepresented in the experiment, which can lead to biased results. For example, if the participants were all recruited from the same university, they may have shared certain characteristics that influenced their behavior.

The adjustment and anchoring bias could have impacted how the participants in the Milgram experiment made decisions. If they were anchored to a particular belief or authority figure, they may not have been able to adjust their behavior or question their actions. For example, if the participants were told that the experiment was important for the advancement of science, they may have felt pressure to continue even if they had doubts about the ethics of the experiment. Confirmation bias can impact how the results of the Milgram experiment were interpreted. If the researchers were biased toward certain outcomes or predictions, they may have interpreted the results in a way that supported their preconceptions. Additionally, if the evaluation metric used to assess the behavior of the participants was biased towards certain outcomes, it may not accurately reflect the ethical implications of the experiment.

Hindsight bias can impact how the behavior of the participants in the Milgram experiment is evaluated. If their actions are judged based on the knowledge we have today, rather than the context in which the experiment was conducted, we may unfairly judge their behavior as unethical. The conjunction fallacy can impact how the results of the Milgram experiment are interpreted. If the researchers overestimate the likelihood of certain events occurring together, they may make incorrect conclusions or recommendations. For example, if the researchers concluded that the behavior of the participants was influenced only by the authority of the experimenter, they may have overlooked other factors that contributed to the outcome.

In conclusion, cognitive biases can impact the selection of participants, interpretation of results, and evaluation of behavior in the Milgram experiment. It is important to be aware of these biases and to take steps to mitigate their impact. By doing so, people can learn from the past and avoid repeating their mistakes.

HOW BIASES AFFECT ALGORITHMS AND THEIR DECISION-MAKING PROCESSES

Cognitive biases can impact the decision-making process, whether it is a human decision or one made by a machine learning algorithm. Biases are errors in judgement that can occur due to a range of factors, including pre-existing beliefs, limited information, or heuristics that humans or algorithms use to simplify complex decision-making processes. In other words, cognitive biases can also manifest in the training and use of supervised learning algorithms, which can lead to inaccuracies and errors in prediction.

The representativeness heuristic can influence how training data is selected for a supervised learning algorithm. If the training data is not representative of the population or the problem being solved, the algorithm may not perform well when applied to new data. For example, if a supervised learning algorithm is trained on data that is biased towards a particular demographic, such as white males, it may not perform well when applied to data from other demographics. Thus, representativeness can lead to stereotypes and biases based on race, gender, or other factors. The availability heuristic can impact the features that are included in a supervised learning algorithm. If certain features are more easily available or accessible, they may be overrepresented in the algorithm, which can lead to inaccurate predictions. For example, if a supervised learning algorithm is trained on data that only includes certain types of crime, it may not perform well when applied to data that includes other types of crime. Availability can lead to overestimating the likelihood of rare events, such as plane crashes, based on media coverage.

The adjustment and anchoring bias can impact how a supervised learning algorithm makes predictions. If the algorithm is anchored to a particular value or parameter, it may not be able to adjust to new information or changes in the problem being solved. For example, if a supervised learning algorithm is trained on data from a particular time period, it may not be able to accurately predict outcomes in the future. Adjustment and anchoring can lead to inaccurate assessments of value, such as when people or algorithms rely too heavily on an initial estimate. Confirmation bias is when individuals or algorithms tend to look for evidence that confirms pre-existing beliefs or hypotheses. This bias can lead to a confirmation of biases that may not be accurate or fair. Confirmation bias can impact how a supervised learning algorithm is trained and evaluated. If the training data is biased towards certain outcomes or predictions, the algorithm may not be able to accurately predict outcomes that do not fit these preconceptions. Additionally, if the evaluation metric used to assess the performance of the algorithm is biased toward certain outcomes, it may not accurately reflect the algorithm's performance in the real world.

Hindsight bias is the tendency to believe, after the fact, that an event was predictable and should have been foreseen. This bias can lead to overconfidence in past decisions or predictions, even if they were based on limited information. Hindsight bias can impact how a supervised learning algorithm is used to make decisions. If the predictions made by the algorithm are overestimated or overconfident, they may be used to justify decisions that are not based on accurate information or evidence. The conjunction fallacy can impact the accuracy of predictions made by a supervised learning algorithm. If the algorithm is overestimating the likelihood of certain events occurring together, it may make incorrect predictions or recommendations. For example, if a supervised learning algorithm is trained to predict the likelihood of a customer purchasing a product, it may overestimate the likelihood if the customer has already purchased a related product, even if there is no actual correlation between the two purchases.

Additionally, biased decisions in sequential ruling occur when the order in which information is presented impacts decision-making. This can affect supervised learning algorithms if they are trained on data that is presented in a particular order. If the algorithm is not designed to account for these biases, it may perpetuate them in its decision-making process.

These processes can lead to errors in judgement and can impact decision-making. To mitigate these biases, it is important to be aware of them and take steps to control their impact. This can include reviewing and selecting unbiased data, analyzing and identifying potential biases in the data, and designing algorithms that are not affected by these biases. This can ensure that the decisions made by both humans and machine learning algorithms are fair, accurate, and unbiased.

WAYS TO CONTROL ALGORITHMIC RESULTS AND INPUT DATA FROM HUMANS

To control the impact of cognitive biases on the results and the training data of supervised learning algorithms, there are several approaches that can be taken. One way to ensure that the data used to train supervised learning algorithms is diverse, representative, and unbiased is to collect data from a variety of sources and to ensure that the data is balanced with respect to different groups and variables. Validating the data by checking for consistency, completeness, and accuracy is also crucial.

Choosing algorithms that are appropriate for the data and the problem being solved is another crucial step in controlling the impact of cognitive biases. This can be achieved by evaluating different algorithms and selecting the one that has the best performance in terms of accuracy and generalization.

Using appropriate evaluation metrics is also important for controlling cognitive biases in supervised learning algorithms. It is crucial to select metrics that are relevant to the problem and that consider different aspects of performance, such as precision, recall, and F1 score. This can help to avoid introducing biases and ensure fair and accurate performance evaluations. Using techniques to detect and correct for bias in the data and the algorithm is also crucial to controlling cognitive biases in supervised learning algorithms (Powers 2007: 1). Analyzing the data for patterns of bias and using techniques such as bias-correction or adversarial training can help to reduce the impact of biases on the algorithm.

Ensuring that teams involved in the development and implementation of supervised learning algorithms are diverse and representative is also an important step. Having teams with different backgrounds, experiences, and perspectives can help to identify and address biases in the data and the algorithm. Also, this is crucial to consider in order to ensure that AGI (artificial general intelligence) systems are not trained on biased data that could perpetuate or even exacerbate societal inequalities.

CONCLUSION

In conclusion, it is important to recognize that controlling cognitive biases in supervised learning algorithms is a complex and ongoing process that requires continuous attention and effort. While there are various approaches that can be taken to mitigate biases, no single method can guarantee completely unbiased and fair results. However, by implementing a multifaceted approach that encompasses data collection, feature selection, algorithm selection, evaluation metrics, bias detection, and diversity in teams, the attempt can be made to minimize the impact of biases and achieve more accurate and equitable outcomes.

Ultimately, it is the responsibility of developers and users of supervised learning algorithms to remain vigilant and dedicated to the pursuit of fairness and objectivity.

REFERENCES

Clarke, Arthur C. *A Space Odyssey*. New American Library, 1968, 2001.

Cialdini, Robert B. *The Psychology of Persuasion*. New York: William Morrow, 1993.

Danziger, Shai, et al. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, vol. 108, no. 17.

Fischhoff, Baruch. Hindsight Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, vol. 1, no. 3.

Jones, Craig E. The Troubling New Science of Legal Persuasion: Heuristics and Biases in Judicial Decision-Making. *Advocates' Quarterly*, 2013, vol. 41, no. 1. Kahneman, Daniel. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

Kahneman, Daniel, et al. Choices, Values, and Frames. *American Psychologist*, 1984, vol. 39, no. 4.

Peer, Eyal, et al. Heuristics and Biases in Judicial Decisions. *Court Review*, 2013, vol. 49, no. 2.

Pereira, Clara Martins. Reviewing the literature on behavioral economics. *Capital Markets Law Journal*, 2016, vol. 11, no. 3.

Powers, David M. W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies,* 2011, vol. 2, no. 1.

Rassin, Eric, et al. Let's Find the Evidence: An Analogue Study of Confirmation Bias in Criminal Investigations. *Journal of Investigative Psychology and Offender Profiling*, 2010, vol. 7.

Sunstein, Cass R. *One Case at a Time: Judicial Minimalism on the Supreme Court.* Cambridge: Harvard University Press, 1999.

Taleb, Nicholas Nassim. *Fooled by Randomness*. New York: Thomson / Texere, 2004.

Tversky, Amos, et al. Judgment under uncertainty: Heuristics and Biases. *Science*, 1974, vol. 185, no. 4157.

Tversky, Amos, et al. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 1983, vol. 90, no. 4.

Wason, Peter. On the Failure to Eliminate Hypotheses in a Conceptual Task. *The Quarterly Journal of Experimental Psychology,* July 1960, vol. 12, no. 3.

Wistrich, Andrew J., et al. Can judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding. *Pennsylvania Law Review*, 2005, vol. 153.

Marcel Piterman
Portugalijos katalikiškojo universiteto Teisės fakultetas,
Katalikiškos teisės ateities tyrimų centras, Portugalija

MAŠININIO MOKYMOSI ŠALIŠKUMO KONTROLĖ: ŽMOGAUS ĮTAKOS MAŽINIMAS PRIIMANT ALGORITMINIUS SPRENDIMUS

SANTRAUKA. Straipsnyje daroma prielaida, kad žmogaus veiklą gali paveikti išorinės aplinkybės, kurios įvardijamos kaip šališkumas. Pažinimo šališkumas susistemintas siekiant nustatyti euristinius procesus, kuriais nesąmoningai mažinamas užduočių sudėtingumas, nors tai gali sukelti sisteminių loginių klaidų. Kaip parodė Milgramo atliktas eksperimentas, žmonės linkę paklusti autoritetui net ir tais atvejais, kai nurodomi veiksmai prieštarauja jų asmeninei sąžinei. Dauguma žmonių visiškai paklusta duotiems nurodymams. Taigi, mašininis mokymasis apima informaciją, kurią žmonės pateikia algoritmams ir kuri gali būti šališka arba turėti prieštaringas instrukcijas, todėl būtina sukurti būdus, kurie padėtų valdyti algoritminius rezultatus ir žmonių iš pradžių pateiktus duomenis.

RAKTAŽODŽIAI: mašininis mokymasis, šališkumas, algoritmai, Milgramo eksperimentas, sprendimų priėmimas.