Artificial Intelligence Decision-Making: Trust Based on Understanding

Chengxiang Wen[1,a], Jiaoyi Wu[2,b,*]

[1]School of Digitalized Intelligence Engineering, Hunan Sany Polytechnic College, Changsha, China

[2]Law School Peking University, Beijing, China

[a]775364962@qq.com, [b]412677236@qq.com

*Corresponding author

Keywords: artificial intelligence; Algorithm discrimination; Algorithm interpretability; Public administration

Abstract: While artificial intelligence has become a key strategic element in many government strategic plans worldwide, there are not many reports and empirical research analyses on the application of artificial intelligence in the field of public administration. This article focuses on analyzing and discussing some issues in the current application of artificial intelligence in the field of public administration, such as accuracy of decision-making, algorithmic fairness, and algorithmic discrimination. What this article attempts to illustrate is that with the development and progress of artificial intelligence, its decision-making will be more reliable than that of humans. Additionally, some questions worth considering and researching are also raised in the article.

1. The application of artificial intelligence in decision making

The life of modern people has unconsciously been deeply embedded in the decision-making system of artificial intelligence: lying on the sofa, the app on the phone [recommendation algorithm] recommends products that users may be interested in based on browsing records. Walking on the street, I passed a autonomous vehicle based on ADABoost; The wildfire situation in the distant forest is monitored by autonomous drones equipped with sensors using deep reinforcement learning algorithms; Below the road surface, the failure of the drainage pipeline is predicted by the stochastic process, and above the road surface, the operation of the power grid is determined by the algorithm in the field of multi-agent system[1]. From business to the public sector, the government's doors are gradually opening up to artificial intelligence decision-making.

On February 19, 2020, Stanford University and New York University jointly released a report to the United States Executive Council titled "Algorithmic Governance: Artificial Intelligence in Federal Administrative Agencies".[2] The report indicates that machine learning and other artificial intelligence technologies are playing an increasingly important role in federal agency rulings, law enforcement, and other regulatory activities. This report summarizes the current use of AI technology in 142 of the most important departments in the United States, providing examples of seven particularly influential/potential application scenarios such as law enforcement, adjudication, and public service provision. It shows that algorithms can help reduce governance costs, improve decision-making quality, and unleash administrative data power. They also consider the challenges of accountability, transparency, and non discrimination in artificial intelligence governance.

As the practice of artificial intelligence shifts from business to the public sector, and governments around the world gradually plan to transition to intelligent governments, some negative concerns about artificial intelligence may be more likely to surface People may not be very sensitive to the potential negative impact of AI decision-making on recommending news, songs, or movies, but when AI can determine public affairs such as government welfare distribution and criminal rulings, it can raise awareness of "robot rule". But the good news is that the current argument is that AI governance is compatible with current laws, and the government's adoption of AI is still slow. [3]

When artificial intelligence has become a key strategic element in many government strategic

  DOI: 10.25236/icetmr.2023.007

plans around the world, China, as the second largest country in the global vitality of artificial intelligence, has a huge amount of data and talent behind it. However, there are currently not many reports and empirical research analyses on the application of artificial intelligence in the field of public administration.[4] There are reports that data-driven governments may actually be more fair, transparent, and responsive than the human face in officialdom, and policy makers use data to provide fast, efficient, and fair public services.[5] This is a very optimistic and positive view on the further application of artificial intelligence. In this regard, the first thing this article aims to illustrate is that artificial intelligence decision-making is more reliable than humans.

2. The accuracy of artificial intelligence decision-making

Although the philosophical reverence for rationality by humans can be traced back to ancient Greece, and modern people in daily life always prefer to consider rationality as their own, psychologists have long discovered that due to the limitations of human cognition and computational abilities, the rationality of decision-makers is limited[6]; Of course, besides not being rational enough, human decision-making tends to take various shortcuts[7], influenced by various emotions[8], and influenced by the social environment[9]. In short, people are not so reliable, and the notion of cognitive biases in decision-making that affect outcomes is not new.

Daniel Kahneman won the Nobel Prize in Economics in 2002 for integrating insights from psychological research into economic science, particularly in human judgment and decision-making under uncertainty.[10] Kahneman is a master figure in the study of human decision-making and judgment, known for his heuristic thinking that studies cognitive biases in uncertain judgments.[11] Not long ago, Kahneman, Sanstan and Siboni proposed in Noise: Defects in Human Judgment that decision makers' judgments are always easy to make mistakes because of different emotions, personalities, pressures, fatigue, cognitive style, and skills. The unnecessary difference in these judgments is "Noise", which may be a relatively new term.[12] The author of 'Noise' believes that decision-making errors are composed of noise and bias, and vividly describes the bias and random distribution errors as systematic biases using target plots. The noise is also classified and statistically expressed.

One example cited in the book regarding the presence of decision noise is the decision of a judge. In criminal justice decisions, temperature, fatigue, hunger, and winning or losing football games can all cause differences in judges' judgments, creating noise in their judgments. The impact of these noises is much greater than we imagine - the differences brought about by different judgments in the same case can lead to endless injustices, high economic costs, and various types of errors. Kahneman does not believe that random error is the noise here (Chapter 17), which is somewhat confusing. In experimental disciplines, the sources of errors include systematic error, random error, and negligent error. The explanations for these three types of error classifications are more in line with my personal cognitive habits, but Kahneman's statement is simple in image and easier to persuade people to demonstrate the reliability of algorithmic decision-making - there are only systematic errors originating from data, models, etc .

What should we do about noise? In order to reduce noise in judgments and improve decision accuracy, Kahneman et al. proposed several principles of "decision hygiene", including using algorithms to replace human judgments, as algorithms have no noise and eliminate a potentially significant source of error in decision-making errors.

One of the meanings in artificial intelligence can be understood as rational behavior, thinking, and doing the right thing.[13] It indeed seems to be in line with the conclusion that decision-making is better than humans, which has been proven by many studies. For example, Jon Kleinberg compared judge decisions and algorithmic predictions in bail decisions and found that machine learning based predictions would bring more benefits to society - reducing crime rates while incarceration rates remain constant, or reducing incarceration rates while crime rates remain constant - due to noise interference in judge decisions.[14] For example, researchers in the Netherlands have shown that the use of algorithms helps decision-makers make more accurate decisions, but there are inherent advantages and disadvantages between algorithms,[15]which will

not be listed one by one.

However, it is worth noting that human thinking is not entirely useless. For example, Kahneman also pointed out that in fields such as literary and artistic creation, the noise of human thinking is precious and worth protecting, so artificial intelligence is not needed to make decisions on behalf of others. Perhaps it can be further clarified that the distinction between accuracy and creativity stems from the difference between artificial intelligence algorithms and human "thinking" methods - the former is a global statistical thinking that is detached from the situation, while the latter is a causal explanation based on the outcome/current situation. Causal thinking allows people to gain a sense of control in uncertainty while ignoring the lack of true precision in noise, but fortunately, it gains a creativity - the limitations of artificial intelligence.

3. The biases of artificial intelligence

3.1 Bias, prejudice, and discrimination

Artificial intelligence decision-making can indeed avoid noise, but deviations are inevitable.

The bias of artificial intelligence (Bias) is not a new concept. Technically, it is often referred to as the inductive bias caused by specific model assumptions[16]. However, in current social discussions, Bias is often explained as the tendency of artificial intelligence systems to make decisions that are considered unfair, support or oppose a person/group[17] - somewhat like bias, but also somewhat like discrimination. In psychology, prejudice is a negative attitude, discrimination is a negative behavior, and the root of discriminatory behavior often lies in prejudice. Racism and gender discrimination can refer to individuals' biased attitudes, discriminatory behaviors, and oppressive institutional practices. Social situations breed and maintain prejudice in various ways, and unequal status generates prejudice, which in turn maintains inequality. Interestingly, how is the problem in social science introduced into machines?

Prejudice is rooted in society, reflected in data, and then learned by algorithms, which is a common saying.[18] It can be understood that prejudice is as ancient as human civilization and deeply rooted in our society; As a part of existing systems and structures, artificial intelligence is not creating biases, but often simply introducing or amplifying social biases. For example, in addition to the inherent biases in calculating assumptions and fitting, artificial intelligence is data-driven and needs to consider that its training data itself contains deep biases (including sensitive attributes and their related attributes or representativeness issues in data sampling, etc.). Algorithm processing of the target itself may also contain implicit discrimination (such as classification algorithms). Some algorithm models may cause biases when used in specific scenarios, and then the algorithm feedback strengthens the biases.[19]

3.2 Example bias

In the Western context, the bias (Bias) often criticized in artificial intelligence driven by big data training and machine learning algorithms is racial discrimination. The classic case of algorithmic discrimination in the United States, the COMPAS system for predicting recidivism, seems to be declaring society, and algorithmic decisions may be more accurate than humans, but they may not be as' correct 'either.[20]

In 2016, the public welfare organization ProPublica published a highly influential report, pointing out that the COMPAS system "discriminates" by labeling black people with high risks, resulting in a higher false positive rate [21] (of course, some also pointed out that COMPAS may also have systemic gender bias [[Melissa Hamilton, The sex algorithm, Behavioral Sciences&the Law, 2019; 37:145-157.][22]). In response, Northpoint (now Equivant), the development company of COMPAS, denied racial discrimination. Northpoint reanalyzed using the same data, but the results showed that COMPAS's predictions for black and white populations were fair, as the recapture rates (or Positive Predictive Values) were the same for different races in high-risk populations. Northpoint believed that ProPublica had made statistical and technical errors.[23] But the debate is far from over, and it can be seen that the reason for the different conclusions lies in the different

standards emphasized by both parties for fair/unfair use - is fairness based on the false alarm rate or the recapture rate? But in fact, these are the evaluation indicators based on confusion matrix for the calculation accuracy in the evaluation classification algorithm; In response to this incident, more analyses have emerged, but they either attempt to reconcile these fair definitions or assert that one or the other is correct - what lies behind this is actually the incompatibility between fair definitions. [24]

3.3 Algorithm fairness

Regarding fairness, it's easy to think that this is not a new issue that artificial intelligence is only facing. Since the 1960s, psychometrists have studied the fairness of educational testing based on their ability to predict performance (in school or at work). [25]At present, the recognition of fairness in computer science is also based on research in other disciplines over the past few decades - education fairness and recruitment fairness are mandatory courses in algorithmic fairness research. [26]A study in 2018 showed that in order to achieve algorithmic fairness,[27] there are more than 20 different definitions of fairness in the literature of computer science, each with its own principles. There are also different opinions on which concept to apply in each scenario, and different concepts have different judgments on whether the same case is fair. The 'Fairness' entry on Wikipedia also indicates that this is already a page full of ambiguity[28]; The concept of fairness constantly changes depending on the country, culture, era, and application field[29]; So, what exactly are we talking about when we're talking about fairness? This article does not define fairness, but only indicates that there may be issues with different standards in the algorithm's attempts to achieve fairness, as this not only affects the judgment of whether there is discrimination in the algorithm, but also affects the algorithm's efforts to eliminate discrimination and achieve algorithm fairness. Many times, it is possible for a single technology to solve a social structural problem, and it is not "fair" to the technology itself.

Algorithm fairness is often discussed in practical applications, such as crime risk assessment and loan issuance, where people focus on the fairness of external decision-making outcomes. The intuitive and simple idea is that in order to avoid racial discrimination, we can try to eliminate racial differences in decision results through simple bias variable elimination, orthogonalization, feature reconstruction, etc.[30]But Kleinberg pointed out that pursuing algorithmic fairness in this way is misleading and the advantages outweigh the disadvantages - regardless of whether sensitive features can truly be removed. He gave an example of the algorithm for determining university admissions, demonstrating that explicitly considering racial characteristics in calculations would yield the most accurate prediction results, while also exhibiting better fairness and efficiency. But then there are questions - such as the appropriate definition and understanding of the concept of fairness, and the ability to set different thresholds for different races to treat each other differently in order to pursue fairness in the outcome. In response, some researchers have proposed new algorithm designs that can maximize public safety while meeting the fairness constraints of reducing racial discrimination, although there is a tension between improving public safety and meeting the popular concept of algorithmic fairness. [31]

In fact, specifically speaking, fairness is reflected within the algorithm as it relates to "selection, assumptions, and definitions".

The implementation of algorithmic fairness is related to the abstraction of many practical issues when designing algorithmic models, such as refining a grand policy vision into a small predictive indicator, simplifying possible complex measures into a limited decision space, and evaluating the predictive performance of the model by selecting appropriate statistical indicators based on the assumption of no interference in a single decision. These abstract operations endow moral and social complex concepts of fairness with mathematical concise expression, which become the characteristics of algorithm runtime V, result Y, score S and decision result D, and are displayed together in the confusion matrix to combine into different definitions of fairness - which creates a huge possible space for achieving fairness in different people's minds.[32]

However, research has shown that the constraints of different definitions of fairness not only

conflict with each other, but also the derived results cannot be achieved in both mathematics and reality. Returning to the previous COMPAS case, if we want to predict results that both (1) the false alarm rate as exemplified by ProPublica and (2) the recapture rate as emphasized by Northpoint are equal across different races, what is needed is the condition known as "utopian" - (1) reality itself is equal and fair (such as no racial difference in the recurrence rate), and (2) the model can accurately predict recidivism (reality cannot be achieved).

Since different forms of fairness cannot be achieved simultaneously, it is necessary to make a choice, which may lead the contradiction back to a classic question of "what to do with fairness". It is only here that algorithms should not be stigmatized for discrimination. It is worth reflecting on the entire society that algorithms provide more solutions to this structural problem. The response of technology to "fairness" is an open issue, and there are many options to try - every step of the algorithm model's selection, assumption, and definition provides the possibility of achieving this fairness. Technicians can think more creatively about policy objectives,make clearer assumptions, clarify trade-offs, and involve the affected public in the entire development process from problem formulation to evaluation, with algorithms subject to regular and meaningful supervision.[33]

4. Ending outlook

This article attempts very limited to illustrate: (1) artificial intelligence algorithms are more accurate in decision-making than human cognition; (2) although artificial intelligence is often criticized for discrimination or bias, the biggest bias often comes from the social reality itself; and (3) artificial intelligence algorithms can achieve fairness in decision-making results by selecting values that align with mainstream fairness. This seems to be a preliminary indication that artificial intelligence should be widely involved in human decision-making and assisted. But in order to make the public truly trust technology, further discussion is about the interpretability/transparency of artificial intelligence that has been widely demanded due to the "algorithmic black box" controversy, which may involve why humans do not trust artificial intelligence technology, whether artificial intelligence technology is truly opaque/unexplainable, what is the interpretable artificial intelligence "XAI" (eXplainable AI) [34], and how to define and "explain" [35], What is the potential negative impact of artificial intelligence algorithm decision-making on society (such as algorithmic monoculture, etc. [36]), and what guidance, supervision, and regulation can the government, society, and technology have in the face of artificial intelligence decision-making.

In short, as long as designed and applied appropriately, especially with the progress of algorithm interpretability, artificial intelligence decision-making can be more accurate, fair, and transparent than humans themselves; There is enormous potential for application in various industries, such as national administrative decision-making. This should be the true transformation and liberation of human thinking and decision-making methods by the so-called information revolution; Each specific step in the process requires social trust based on understanding.

References

[1] Mykel J. Kochenderfer, Tim A. Wheeler, Kyle H. Wray, Algorithms for Decision Making, The MIT Press, 2022, p. 1-16.

[2] David Freeman Engstrom, et al., Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies, NYU School of Law, Public Law Research Paper, 2020 (20-54)

[3] Cary Coglianese, David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, The Georgetown Law Journal, 2016, Vol:105, p. 1147- 1223.

[4] Zhang, Weidong, et al., Factors influencing the use of artificial intelligence in government: Evidence from China, Technology in Society,2021, Vol 66: 101675, p.1-16.

[5] Helen Margetts, Cosmina Dorobantu, Rethink government with AI, Nature, 2019-4-11, Vol 568, p. 163-165

[6] Herbert A. Simon, Bounded rationality, Utility and probability, Palgrave Macmillan, London, 1990. 15-18.

[7] Daniel Kahneman, Shane Frederick, Representativeness revisited: Attribute substitution in intuitive judgment, Heuristics and biases: The psychology of intuitive judgment, Cambridge University Pres, p. 49-81 (2002).

[8] Hans-Rüdiger Pfister, &Gisela Böhm, The multiplicity of emotions: A framework of emotional functions in decision making, Judgment and Decision Making, Vol 3:1, p.5–17(2008).

[9] Xiao TianWang, Frédéric Simons, Serge Brédart,Social cues and verbal framing in risky choice, Journal of Behavioral Decision Making,Vol 14:1, p.1-15 (2001).

[10] The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2002. NobelPrize.org. Nobel Prize Outreach AB 2022. Thu. 5 May 2022, https://www.nobelprize.org/prizes/economic-sciences/2002/summary/

[11] AmosTversky, Daniel Kahneman, Judgment under uncertainty: Heuristics and biases. Science, Vol 185(4157), p. 1124–1131 (1974).

[12] Michael Blastland, Signal failure: Daniel Kahneman's fascinating—and flawed—new book 'Noise', Prospect, 2021-6-24, https://www.prospectmagazine.co.uk/arts-and-books/signal-failure-daniel-kahnemans-fascinating-and-flawed-new-book-noise

[13] Stuart Russel, Peter Norvig, Artificial intelligence: a modern approach, Fourth Edition, Pearson Education, (2021), p. 1-5.

[14] Jon Kleinberg, et al, Human decisions and machine predictions, The quarterly journal of economics, Vol 133:1, p. 237-293(2018).

[15] Marijn Janssen, et al. Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-Supported Decision-Making in Government, Social Science Computer Review, vol. 40, no. 2, Apr. 2022, p. 478–493.

[16] David Haussler, Quantifying inductive bias: AI learning algorithms and Valiant's learning framework, Artificial intelligence,Vol 36: 2, p. 177-221(1988).

[17] Eirini Ntoutsi, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, Vol 10:e1356, p 1-14, 2020.

[18] Solon Barocas, Andrew D. Selbst, Big data's disparate impact, California Law Review, Vol 104,p.671-732 (2016).

[19] Bruno Lepri, et al, Fair, transparent, and accountable algorithmic decision-making processes, Philosophy & Technology, Vol 31: 4, p. 611-627(2018).

[20] Sam Corbett-Davies, et al, A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear, The Washington Post, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10 /17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

[21] Julia Angwin, et al., Machine Bias, May 23, 2016, ProPublica., https://www. propublica. org/article/machine-bias-risk- assessments-in-criminal-sentencing.

[22] Melissa Hamilton, The sexist algorithm, Behavioral Sciences & the Law, 2019; 37: 145–157.

[23] William Dieterich et al., COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, NORTHPOINTEINC. Research. Department'T, July 8, 2016, http://go. volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

[24] Shira Mitchell, et al., Algorithmic fairness: Choices, assumptions, and definitions, Annual Review of Statistics and Its Application,Vol 8, p. 141-163(2021).

[25] Richard B. Darlington, Another Look At “Cultural Fairness”, Journal of Educational Measurement, Vol 8:2, p. 71-82 (1971).

[26] Ben Hutchinson, Margaret Mitchell, 50 years of test (un)fairness: Lessons for machine learning, Proceedings of the Conference on Fairness, Accountability, and Transparency, p. 49-58 (2019).

[27] Sahil Verma, Julia Rubin, Fairness Definitions Explained, 2018 IEEE/ACMInternational Workshop on Software Fairness (Fairware), p. 1-7 (2018).

[28] https://en.wikipedia.org/wiki/Fairness

[29] Marie Schäfer, Daniel B. M. Haun, Michael Tomasello,Fair Is Not Fair Everywhere, Psychological Science, Vol 26:8, p.1252-1260 (2015).

[30] Jon Kleinberg, et al., Algorithmic Fairness, Aea papers and proceedings, Vol: 108, p. 22-27 (2018).

[31] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness, In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 797–806.

[32] Supra note 28, Shira Mitchell: p. 148-150.

[33] Supra note 28, Shira Mitchell: p. 157-159.

[34] Gunning, David, et al. XAI—Explainable artificial intelligence, Science Robotics Vol 4:37, eaay 7120 (2019).

[35] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence, Vol. 267, p.1-38 (2019).

[36] Jon Kleinberg, Manish Raghavan, Algorithmic monoculture and social welfare, Proceedings of the National Academy of Sciences Vol. 118: 22, p. 1-35 (2021).