Catastrophe and Climate

Risks Emerging from Artificial Intelligence Widespread Use: A Collection of Essays

AUGUST | 2024

SOA Research INSTITUTE

SOA Research INSTITUTE

Risks Emerging from Artificial Intelligence Widespread Use

A Collection of Essays

CONTENTS

Risks Emerging from Artificial Intelligence Widespread Use

A Collection of Essays

Introduction

INTRODUCTION

The Society of Actuaries Catastrophe & Climate Strategic Research Program Steering Committee issued a call for essays to gather Risks Emerging from Artificial Intelligence Widespread Use. The purpose was to gather potentially catastrophic uses of artificial intelligence (AI) as submitted by essay authors with the goal to initiate discussion regarding risks and impact of AI and set the stage for future research.

The six essays that form this collection are included below. The top two essays chosen for creativity, originality and the extent to which an idea might help promote further thought in this area, are noted here:

First Prize	**Risks Emerging from Artificial Intelligence Widespread Use** Hanchen (Henry) Wang and Yongqi Liang
Second Prize	**In Praise of Actuarial Judgment and the Dangers of Relying on Historical Data** Sam Gutterman, FSA, CERA, FCAS, MAAA, FCA, HonFIA

Thank you for your interest in this essay collection. We welcome your feedback via the survey banners embedded in this document.

THE CALL FOR ESSAYS

BACKGROUND AND OVERVIEW

The capacity and availability of artificial intelligence (AI) have expanded dramatically in recent years. What was once the domain of major information technology companies and specialists is now accessible to individuals, small businesses, and public service providers. However, as AI becomes more pervasive, we must consider the potential repercussions and risks associated with its widespread use.

The Society of Actuaries Research Institute (SOA) Catastrophe & Climate Strategic Research Program seeks essays that delve into the multifaceted risks posed by (AI) over both the short and long term. Accepted essays will be published on the SOA webpage and will explore potential risks associated with a large-scale use of AI. In particular, we are looking for discussion of one or more potentially catastrophic risks that could result from unintended or misuse of AI.

With these issues in mind, the Society of Actuaries Research Institute is interested in an exploration of this topic from a variety of perspectives. The result of this effort is intended to initiate discussion regarding risks and impact of AI and set the stage for future research.

As Artificial Intelligence (AI) continues to advance, it holds immense promise for improving human lives like AI-powered medical diagnoses or more efficient traffic management. However, alongside its potential benefits, there are also significant risks that could emerge from its use. These are several important overarching considerations regarding the scope of this call for essays:

Short-Term AI Risks:

- Magnitude of Consequences: In the short term, AI risks may manifest as privacy breaches, algorithmic bias, and job displacement. These immediate consequences impact individuals, organizations, and society.
- Quantification: Quantifying these risks involves assessing the likelihood of specific events (e.g., a data breach) and their potential impact (e.g., financial losses, reputational damage).
- Costs and Uncertainties: Organizations must weigh the costs of implementing safeguards against the uncertainties of risk occurrence.

Long-Term AI Risks:

- Existential Threats: Over the long term, AI could pose existential risks. These include scenarios where AI surpasses human control, leading to unintended consequences.
- Balancing Priorities: Balancing short-term benefits with long-term risks requires thoughtful consideration.
- AI Alignment: Ensuring AI systems align with human values is crucial. Misaligned AI could lead to catastrophic outcomes.

ESSAY CONTENTS

The organizers seek essays that address potentially catastrophic risks emerging from AI widespread use. This invitation has been deliberately written broadly to allow respondents the flexibility to address this topic from one or more perspectives and approaches. Respondents are free to choose from one or more of the following sample topics below or propose others that fall within the scope of this area as described in the above section. Please note that the list is not meant to be exhaustive but merely examples of proposed topics that may be considered. Respondents are welcome to address other questions or topics that fall under the general scope of this call for essays.

Sample topics include:

- Black box AI: Lack of transparency in ai decision-making can compromise accountability and trust.
- Privacy Erosion: AI systems handling personal data could compromise privacy if not properly regulated.
- Economic Disruption: Widespread automation due to ai could lead to job losses, economic inequality, errors in social benefit entitlements and social instability.
- Biased Decision-making: AI algorithms trained on biased data may perpetuate discrimination and exacerbate societal inequalities.
- Malicious Use: AI can be weaponized for cyberattacks, misinformation, and surveillance.
- Geopolitical Risks associated with AI.
- The impact on insurance coverage or need for new products to address these risks.

First Prize Winner

Risks Emerging from Artificial Intelligence Widespread Use

Hanchen (Henry) Wang and Yongqi Liang

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

ABSTRACT

This paper explores the risks posed by widespread artificial intelligence (AI) use on individual privacy and proposes a framework for integrating data into decision-making with a focus on fairness, accountability, and transparency. Through real-world examples and analysis, we examine how AI technologies can compromise privacy and highlight key challenges such as biases and the need for informed consent. The paper highlights the urgency for regulatory and ethical interventions to address these issues. Our framework is designed to support organizations in making responsible ethics-based data decisions with a priority on maintaining societal well-being. By proposing frameworks to preserve privacy while fostering AI innovation and promoting ethical data practices, this research contributes to the discourse on responsible AI governance and fosters data use safety within diverse organizations.

INTRODUCTION

The rapid development and widespread adoption of AI have transformed and promised significant benefits to various sectors from healthcare to finance. However, this transformation brings with it a wide range of risks that need careful consideration. This essay will explore the potential short-term and long-term risks associated with AI's widespread use. We will focus on privacy erosion, economic disruption, and biased decision-making, along with the broader existential threats that AI might pose.

SHORT-TERM RISKS

PRIVACY EROSION

AI systems rely heavily on vast amounts of data to function effectively. This often includes personal and sensitive information. The potential for data breaches and misuse of personal data is high, especially if AI systems are not adequately regulated. Privacy breaches can lead to significant financial losses, reputational damage, and a loss of trust among users.

Examples:

- *Camera Surveillance in China*: The extensive use of AI-powered surveillance cameras in China for monitoring citizens has raised significant privacy concerns. These cameras, often equipped with facial recognition technology, are deployed in public spaces to track individuals' movements and behaviors. This level of surveillance can lead to invasions of personal freedoms and privacy. This is due to the fact individuals may feel constantly watched and monitored by the state. The potential misuse of this data for political or social control further exacerbates the ethical issues surrounding AI surveillance (Yang, 2022). The pandemic has further accelerated the use of such technologies, justifying them under the guise of public health and safety. This cover-up has led to widespread acceptance among the population (Yang, 2022).
- *Human Tracking Devices in Self-Driving Cars*: The use of AI to track and collect data on individuals' movements and behaviors through human tracking devices in self-driving cars is another area of concern. While these devices can enhance navigation and safety features, they also collect vast amounts of personal data. U.S. corporations might even collect massive amounts of data on individuals to sell to the Chinese government for profit. The Chinese government even installed an unauthorized surveillance camera inside the Tesla Auto-Car made in China to track down individuals privately and monitor their daily activities (Schellekens, 2022).
- *Amazon Just Walk Out Technology*: Amazon's Just Walk Out technology allows customers to shop without going through the traditional checkout processes. They do this by using AI to track items picked up and placed down to automatically charge their accounts. While convenient, this technology raises privacy concerns as it involves constant monitoring of customers' shopping behaviors. The data collected can provide detailed insights into individuals' purchasing habits. This data could be used for targeted advertising and manipulation of consumer choices. This seamless shopping experience, while revolutionary, presents significant privacy challenges as every item selected by a customer is tracked and recorded. This detailed consumer profile could be exploited for commercial purposes without the consent of the consumers. In addition, Amazon would need its customers to accept terms and conditions to access its service. Once customers click the "accept" button, they grant Amazon extensive rights to their data, which allows Amazon to legally collect and invade customers' shopping preferences. Customers lose their privileges to demand justice once Amazon causes conflicts of interest with customers. Terms and Conditions clear Amazon when customers want to settle a lawsuit against them for violating their "privileged rights" (Ives et al., 2019).

ECONOMIC DISRUPTION

AI-driven automation has the potential to disrupt labor markets significantly. While AI can enhance productivity and efficiency, it also poses a risk of job displacement. This could lead to increased economic inequality and social instability. The transition to an AI-driven economy requires careful management to ensure that the benefits of AI are distributed equitably and that displaced workers are supported through retraining and social safety nets.

Examples:

- *Social Media Tracking*: Platforms like Instagram and TikTok use AI to track user behavior can potentially infringe on privacy and contribute to data misuse. AI algorithms analyze user interactions to personalize content, advertisements, and recommendations. While this enhances user experience, it also means that vast amounts of personal data are being collected and analyzed without users' explicit consent. The U.S. banned TikTok and integrated TikTok features in Instagram because any corporations outside of China can turn over all data directly to the Chinese Communist Party. The main issue is access to individuals' entire phones: search history, microphone, geolocation, metadata, and camera. The Chinese government is more powerful than any multinational corporation in the mainland of China. This means they can use their political power to exploit all user data from other countries. This can lead to privacy breaches and exploitation of user data for the Chinese Communist Party's commercial gains. On the other hand, U.S. corporations are more powerful than the U.S. government, which indicates that social media platform users' data can be taken by U.S. corporations and sold to the U.S. government as well. The U.S. government could use data to track down individuals to access their records so they can find a way to change their values and beliefs. Many U.S. social media corporations will simply vote for those who push policies that benefit them the most (Perakakis et al., 2019). Ultimately, U.S. corporations other than social media companies use data to target audiences with AI. Companies leverage AI to analyze consumer data and tailor marketing strategies. However, oftentimes these companies do so without clear transparency or consent from users. This practice raises ethical concerns about data ownership, consent, and the potential for manipulative marketing tactics that exploit personal information (Svetlana et al., 2022).

BIASED DECISION-MAKING

AI algorithms are only as good as the data they are trained on. If the training data is biased, the AI system will likely perpetuate these biases, leading to discriminatory outcomes. This is particularly concerning in critical areas such as hiring, law enforcement, and access to financial services. Addressing algorithmic bias is essential to prevent AI from exacerbating existing societal inequalities.

Examples:

- *Education*: The use of AI in education, particularly tools like Zoom during the pandemic, highlights the potential dangers and changes in learning dynamics. While AI-powered platforms facilitated remote learning, they also introduced challenges such as unequal access to technology and data privacy concerns. Additionally, AI algorithms used for grading and student assessments can perpetuate biases and deprive certain groups of students based on their socioeconomic background or learning style. For example, Slimi and Carballido highlight that students with different skin colors struggle more academically due to their socioeconomic status, and women of color face significant barriers in STEM fields due to social isolation and biases (Slimi & Villarejo Carballido, 2023).
- *Online Proctoring Software*: Yoder-Himes conducted research on the bias in facial recognition technology used in online proctoring software to detect cheating. The study revealed that students with darker skin tones were more likely to be flagged for cheating, leading to increased scrutiny from instructors. This bias against students based on race and gender highlights the ethical concerns surrounding the use of AI in educational settings and the need for further research to understand and mitigate algorithmic biases (Yoder-Himes et al., 2022).
- *Use of ChatGPT*: The use of AI like ChatGPT may influence tendencies among university students. Muhammad, Ahmed, and Tariq conducted research on the causes and effects of general use of ChatGPT among university students. They compared and contrasted those who frequently use ChatGPT and those who do not to see if ChatGPT influences academic tendencies among university students. Their findings also suggest that students whose academic workload and time pressure is higher report higher frequent use of ChatGPT to cope with their stressful academic circumstances. Their findings also suggest that those who are sensitive to academic rewards report lower use of ChatGPT than those who aren't. Furthermore, excessive use of ChatGPT can cause mental damage to students and their academic outcomes. Those who always use ChatGPT tend to procrastinate their work and report memory loss compared to those who barely use ChatGPT. This is because students who frequently rely on ChatGPT lose their critical thinking and problem-solving skills, which can harm their academic performance. It can be concluded that ChatGPT usage correlates with time pressure, academic workload, and sensitivity to rewards with students' academic outcomes (Abbas et al., 2024).

LONG-TERM RISKS

EXISTENTIAL THREATS

In the long term, AI could pose existential risks if it surpasses human control. The development of superintelligent AI systems could lead to scenarios where AI operates beyond human oversight, making decisions that could have catastrophic consequences. Ensuring that AI systems are aligned with human values and goals is crucial to mitigating these risks.

Examples:

- *Tesla Driverless Cars*: Recent news about driverless cars and potential data breaches or hacking could lead to uncontrolled scenarios. If AI systems in autonomous vehicles are compromised, it could result in widespread chaos and accidents. The potential for hackers to take control of driverless cars poses significant risks to public safety and highlights the need for robust cybersecurity measures in AI systems. The safety argument emphasizes the benefits of AI in driving accuracy but also highlights the need for extensive testing and reliable data to ensure safety (Blanco et al., 2016).
- *Statistical Data*: "The Safety Argument" describes the decisional phenomenon of autonomous vehicles (AV) by comparing and contrasting AI decisional capacity with real human drivers. Statistical driving data highlights the correlation between erroneous human driving decisions and road accident fatalities. The safety argument of AV technologies stresses the core safety benefits that can promote more accurate driving abilities and advocate advanced AI decision capacity to accurately navigate the road network. However, research studies such as from Blanco illustrate that by comparing and contrasting AV driving data to human driving data "The Safety Argument" cannot accurately determine the outcome of the AI decisions (Blanco et al., 2016). This is because "The Safety Argument" cannot be predicted and justified by data analysis in the AV system. The Virginia Tech Transportation Institute criticized the practice of using inaccurate data for safety analysis, claiming that "with a Poisson distribution and national mileage and crash estimates, automated vehicles would need to drive 725,000 miles on representative roadways without incident and without human assistance to say with 99% confidence they crash less frequently than vehicles with human drivers" (Goodall, 2014a). It seems that inaccurate statistical methods resulting in invalid data outcomes is why an AV occasionally goes wrong and presents different formats and critiques of AV crash report data. Furthermore, sensor error, programming bugs, unanticipated objects, classification errors, and hardware/software faults present further unsolved challenges to the AV safety argument in the future due to the statistical issue of data deficiencies. Blanco claims that future AI technologies cannot guarantee the safety of AV due to higher risks of crashes, stating, "The limited exposure of the self-driving car project to real-world driving increases statistical uncertainty in its crash rate. That uncertainty will decrease as it receives more on-road, in-traffic testing" (Blanco et al., 2016).

AI ALIGNMENT

One of the most significant challenges in AI development is ensuring that AI systems align with human values. Misaligned AI systems could make decisions that are harmful or unintended. Research in AI alignment aims to create systems that understand and prioritize human values, reducing the risk of catastrophic outcomes.

Examples:

- *Healthcare Technology*: AI's impact on healthcare and potential risks associated with data handling and privacy. AI applications in healthcare, such as diagnostic tools and personalized treatment plans, rely on sensitive patient data. Ensuring data privacy and security is paramount to prevent misuse and maintain patient trust. Additionally, AI systems must be designed to align with ethical principles in healthcare, prioritizing patient welfare and informed consent (Rahman et al., 2024).
- *Financial Systems*: AI's role in managing money on the internet, including banks, cryptocurrencies, and NFTs, with potential vulnerabilities to data breaches. The financial sector's increasing reliance on AI for transactions, fraud detection, and investment strategies necessitates stringent security measures to protect against cyber threats. The potential for AI-driven financial systems to be exploited for criminal activities underscores the need for regulatory oversight and robust cybersecurity frameworks (Hidayat et al., 2024).

GEOPOLITICAL RISKS

The widespread use of AI also has geopolitical implications. Nations that lead in AI development could gain significant strategic advantages, potentially leading to imbalances in global power dynamics. AI could be weaponized for cyberattacks, misinformation campaigns, and surveillance, all of which pose risks to national security and global stability. "Sub-conscious and personalized levels of algorithmic persuasion may have significant effects on the cognitive autonomy of individuals and their right to form opinions and take independent decisions" (Ashraf, 2021).

Examples:

- *Algorithmic censorship*: The rise of AI technology can threaten freedom of speech online. AI technologies can extract data to deliver customized content and influence users' thoughts and perceptions. This "algorithmic persuasion" can affect individuals' cognitive autonomy and their right to form independent opinions. The integration of AI in business models of "surveillance capitalism" can punish dissenters for not fitting into their own agenda. Zeynep Tufekci's concept of "algorithmic censorship" highlights how AI determines what can be seen online, shaping individuals' online environments, and potentially leading to biased content moderation (Ashraf, 2021).
- *Facebook-Cambridge Analytica Scandal Case Study:* The corporation Cambridge Analytica in 2018 harvested millions of Facebook users' private information and data without their consent for political campaign advertisements. The scandal underscores the ethical dilemmas surrounding data privacy and causes many controversies related to the ethical responsibilities of technology companies in safeguarding individuals' privacy rights. Algorithmic bias in criminal justice illustrates that all algorithmic tools utilized in criminal justice systems have been scrutinized for perpetuating discrimination against marginalized communities. Studies have concluded that algorithms in the criminal justice system could exhibit racial discrimination based on individuals' socioeconomic status in society. This could lead to inequalities in the legal system and disparities in sentencing outcomes. Facebook CEO Mark Zuckerberg acknowledged his failure to protect individuals' data and implemented measures to upgrade data protection and cybersecurity systems (Trout, 2016). Researchers, regulators, and policymakers have demanded greater accountability and fairness in algorithmic decision-making in response to the scandal and concerns about algorithmic bias in the criminal justice system. Additionally, some jurisdictions have implemented bans and restrictions on the use of biased algorithms in the criminal justice system to enhance legal decision-making processes and reduce the risk of perpetuating discrimination. The Facebook-Cambridge Analytica Scandal can teach us that AI can play a vital role in shaping the future of the criminal justice system (Trout, 2016).
- *AI Content Moderation*: Online platforms have turned to AI for better content moderation. AI content moderation is responsible for reporting any negative content by removing offensive wordings, filtering inappropriate comments, and permanently deleting spam. This shows that content moderation generated by AI accumulates massive amounts of user data and applies data science tricks/techniques to identify certain patterns and correlations to hypothesize trends and outcomes to govern speech. Thus, AI can influence the way speech structures are structured by modifying content before being officially published. The use of AI content moderation illustrates biased stereotypes, racial discrimination, and hate speech online. This is because AI content moderation tools automatically target certain vulnerable minority groups and mark them as dangerous threats. AI hate speech detection tools demonstrate biases against African Americans. These systems are more likely to ban their tweets online due to their potential likelihood of physical violence demonstrations and verbal assaults online. AI technologies, like a double-edged sword, can create many opportunities and challenges for us. It is our responsibility to control it and fulfill its potential to benefit us in many aspects of our lives (Ashraf, 2021).

CONCLUSION

The rapid advancement of AI presents both opportunities and risks. While AI holds immense potential to improve various aspects of human life, it is crucial to address the associated risks proactively. Privacy erosion, economic disruption, biased decision-making, and existential threats are significant concerns that need careful consideration and regulation. By understanding and mitigating these risks, we can harness the benefits of AI while minimizing its potential harms. The Society of Actuaries Research Institute's call for this topic is a valuable initiative to stimulate discussion and promote further research in this critical area.

* * * * *

Hanchen (Henry) Wang graduated from the University of California, San Diego with degrees in Business Psychology and International Studies with focus on Economics. He enjoys collecting books, stamps, and different currencies. He enjoys working with data and inspires to become a licensed actuary. He can be reached at 1996920@gmail.com.

Yongqi Liang is an undergraduate student at the University of California, San Diego majoring in Math-Econ with a minor in Finance. He is passionate about actuarial science and is working towards becoming a licensed actuary. He can be reached at yongqiliang2004@gmail.com.

REFERENCES

Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21(1). https://doi.org/10.1186/s41239-024-00444-7

Ashraf, C. (2021). Exploring the impacts of artificial intelligence on freedom of religion or belief online. *The International Journal of Human Rights*, 26(5), 757–791. https://doi.org/10.1080/13642987.2021.1968376

Blanco, M., Atwood, J., Russell, S. M., Trimble, T. E., McClafferty, J. A., & Perez, M. A. (2016, January 8). Automated Vehicle Crash Rate Comparison Using Naturalistic Data. *VTechWorks Repository*. https://vtechworks.lib.vt.edu/items/d33dc4e1-58f3-4813-9677-4a0f425880c7

Goodall, N. J. (2014a). Ethical Decision Making during Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2424(1), 58–65. https://doi.org/10.3141/2424-07

Goodall, N. J. (2014b). Machine Ethics and Automated Vehicles. *Road Vehicle Automation*, 93–102. https://doi.org/10.1007/978-3-319-05990-7_9

Hidayat, M., Defitri, S. Y., & Hilman, H. (2024). The Impact of Artificial Intelligence (AI) on Financial Management. *Management Studies and Business Journal (PRODUCTIVITY)*, *1*(1), 123–129. https://doi.org/10.62207/s298rx18

Ives, B., Cossick, K., & Adams, D. (2019). Amazon Go: Disrupting retail? *Journal of Information Technology Teaching Cases*, *9*(1), 2–12. https://doi.org/10.1177/2043886918819092

Perakakis, E., Mastorakis, G., & Kopanakis, I. (2019). Social Media Monitoring: An Innovative Intelligent Approach. *Designs*, *3*(2), 24. https://doi.org/10.3390/designs3020024

Rahman, Md. A., Victoros, E., Ernest, J., Davis, R., Shanjana, Y., & Islam, Md. R. (2024). Impact of Artificial Intelligence (AI) Technology in Healthcare Sector: A Critical Evaluation of Both Sides of the Coin. *Clinical Pathology*, 17. https://doi.org/10.1177/2632010x241226887

Schellekens, M. (2022). Human–machine interaction in self-driving vehicles: a perspective on product liability. *International Journal of Law and Information Technology*, *30*(2), 233–248. https://doi.org/10.1093/ijlit/eaac010

Slimi, Z., & Villarejo Carballido, B. (2023). Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies. *TEM Journal*, 590–602. https://doi.org/10.18421/tem122-02

Svetlana, N., Anna, N., Svetlana, M., Tatiana, G., & Olga, M. (2022). Artificial intelligence as a driver of business process transformation. *Procedia Computer Science*, *213*, 276–284. https://doi.org/10.1016/j.procs.2022.11.067

Trout, K. E. (2016). The Impact of Electronic Health Records on Healthcare Service Delivery, Patient Safety, and Quality. *DigitalCommons@UNMC*. https://digitalcommons.unmc.edu/etd/173/

Yang, Z. (2022, October 11). The Chinese surveillance state proves that the idea of privacy is more “malleable” than you’d expect. *MIT Technology Review*. https://www.technologyreview.com/2022/10/10/1060982/china-pandemic-cameras-surveillance-state-book/

Yoder-Himes, D. R., Asif, A., Kinney, K., Brandt, T. J., Cecil, R. E., Himes, P. R., Cashon, C., Hopp, R. M., & Ross, E. (2022). Racial, skin tone, and sex disparities in automated proctoring software. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.881449

Second Prize Winner

In Praise of Actuarial Judgment: The Dangers of Relying on Historical Data

Sam Gutterman, FSA, CERA, FCAS, MAAA, FCA, HonFIA

The thesis of this essay is that over-reliance on looking backward, whether on mortality, morbidity, other actuarial practice-related information/data, or artificial intelligence (AI) training data or information, can be dangerous and can lead to sub-optimal findings and recommendations. Despite the broad potential of AI, supplementary actuarial judgment is necessary for the sound application of AI-generated modeling or findings in an actuarial application.

An initial AI Safety Summit in November 2023 produced the Bletchley Agreement. This resulted in a declaration, signed by the European Union and twenty-eight countries, including China, Saudi Arabia, and the United States, which called on signatories to ensure that AI is "human-centric, trustworthy and responsible." I will focus on all three desirable of these aspects.

AI is about to become even more ubiquitous – with overwhelming attention, it will be the dominant theme of numerous conferences and efforts worldwide in all sorts of applications, as well as a primary driver of a booming stock market.

But before I explore the main theme of this paper, it is worthwhile to address the definition of AI, of which several have been put forth. One that I have seen is that it seeks to mimic human intelligence by inferring patterns and connections in data that are not easily discernible to humans. AI systems can certainly analyze large amounts of data faster than we can by ourselves. They can also develop predictions about future outcomes based on historical information.

Given this definition, let's contrast this with actuarial judgment. But there are also similarities. Pattern recognition is common to both and is an underlying contribution to the effectiveness of many actuaries. But that is only part of actuarial analysis, which can provide timely identification of significant problems or issues (which may depend on the application), as well as quantify the effects of key risks/opportunities and approaches to address them.

Too often, actuaries have been characterized as relying solely on historical experience and information – but this can also be a criticism of AI. Both rely on historical experience or information (i.e., referred to as training data in AI), based on a deep analysis of patterns in available data applied in a systematic and algorithmic manner. What may be different from the AI repertoire is that a skilled actuary also applies judgment to the problem at hand.

AI algorithms can develop and present predictions by correlating information and applying patterns from past events or adjustments based on historical data and past learnings. AI models can process vast amounts of historical data quickly, organizing them by rules and using labels ultimately provided by humans, possibly containing an inherent element of bias, although an AI (human) supervisor can train the AI model to identify and thus ignore this bias, at least to the extent provided by the supervisor.

But some AI models can be black boxes — they may not adequately explain to users how they arrived at their conclusions; nor why their insights should or should not be relied upon. If they are wrong, a danger is that AI systems will undermine the trust given to them and in some cases social trust, by leading their users astray.

Although constraints and considerations could be built into an AI model's algorithm, there are certain factors, such as race or ethnicity, that would be unacceptable to use in a function such as insurance pricing but may be acceptable for use in a public policymaking process.

THE DATA, THE PROBLEM, AND THE CONTEXT

Identification and assessment of reliable and relevant data/information input are crucial actuarial functions. A great deal of experience analysis units inside an insurance company, for example, are involved in this sometimes time-consuming, tedious, yet complex process. Hidden problems in experience data exist, just as they can affect the data used in training an AI model. Once reliable and relevant data/information are obtained, analysis can begin.

AI can't by itself ask all the right questions and exhibit skepticism, although they can be applied to a preconceived problem or issue. AI may not, by itself, be able to identify the quality of or bias in the data obtained or to appropriately weight data/information obtained from different sources. It could, by applying an algorithmic rule, assign a degree of credibility to sets of data based on size; but it may not be able to assess reliability or relevance, which may only be applicable to the issue being addressed.

For example, a common problem is dealing with an outlier, which may prove to be an early warning indicator or an indication of faulty data that might be safely ignored. I asked ChatGPT for accepted approaches to assess outliers. What it 'told' me is a decent start in analysis: it depends on the context and purpose of the analysis. It then proceeded to provide some common approaches to use:

1. "Understanding the cause, that is, whether due to measurement error or natural variation;
2. data cleaning;
3. transformation or normalization;
4. statistical methods such as regression or median-based measures;
5. segregation if genuine or imputation if erroneous, that is, analyzing them separately or substituting values surrounding them;
6. modeling; and
7. reporting (transparently) and justifying decisions made in light of the outliers."

It then indicated that the choice of the method used "should be guided by a combination of statistical rigor, domain knowledge, and the specific goals of your analysis." This is certainly (and surprisingly) good advice. But it leaves the choice as to which approach to use to the analyst (who has domain knowledge) – which is the role of actuarial judgment. Also, it suggested that the person (people) who is overseeing the use of an AI tool should ensure that the data/information used is not misinterpreted, is contextually applicable, and whether one or multiple approaches should be applied sequentially or at the same time.

This reminds me somewhat about the time that I over-fitted a set of data. I was so uber-focused on past experience, that is, I slavishly followed patterns, trends, and covariances derived from the past, when subsequent changes in underlying conditions or populations over the period studied would have led me to quite different conclusions. It is easy to lose sight of the objectives of an analysis, which should include an assessment of future risks that should have been anticipated and acting dynamically in response.

On the other hand, the use of the output of an AI model might encourage the user to think more in-depth regarding where to be skeptical, what questions to ask, and what explicit assumptions to make.

The context is important in many actuarial analyses, e.g., in regular monitoring, pricing, valuation, liability analysis, and risk management. Different analytical approaches may be appropriate for each. One model, whether or not an AI one, cannot apply to all applications. The user can provide an AI model with a description of what to do, and some hints (leading language) about how to approach the analysis. Just as the use of a computer is quite a bit more efficient than manually prepared spreadsheets, actuarial-specific software that is regularly used can be even more efficient and useful. Incorporating AI features can, for example, create alternative pre-defined or AI-generated scenarios faster than a new set of scenario-specific assumptions in non-AI software that would have to be created from scratch each time applied.

Many actuarial exercises are predictable – conducted once a year, quarter, or month when an analysis of new data is combined with previous data to form new patterns to use as input into a pricing, funding, or liability calculation. However, that said, I would hate to present to a CEO results that have not been properly supervised, governed (e.g., through a peer review or multi-disciplinary governance committee), and validated.

CHANGES IN CONDITIONS

Training a model or its assumptions on historical data/information can be quite valuable. Unless, of course, the data/information used directly or for training purposes is not relevant to what that experience will be applied to. Certainly, the job of an actuary is often to assess or project the future, but it may be just as if not more important to select the most appropriate historical experience to serve as its basis and to determine the types of adjustments are needed to derive a reasonable estimate of future experience. This places a premium on estimating the extent future risk factors will be different from those of the past.

Because actuarial analysis tends to focus on such risks, I asked ChatGPT what the most significant upcoming mortality risks are. It indicated:

1. "Health-related risks.
    a. Non-communicable diseases: Conditions such as heart disease, cancer, and diabetes continue to be leading causes of death globally. Heart disease remains the top killer, followed by cancer and diabetes-related complications.
    b. Pandemics and infectious diseases: The threat of global pandemics, including influenza and emerging high-threat pathogens like Ebola remains significant. The COVID-19 pandemic has highlighted the devastating impact of infectious diseases on global mortality.
    c. Antimicrobial resistance: This is a growing concern as it can lead to common infections becoming untreatable, resulting in higher mortality rates from previously manageable diseases.
2. Climate-related risks.
    a. Extreme weather events: Increasingly severe weather events such as hurricanes, floods, and heatwaves pose significant mortality risks. These events are expected to become more frequent and intense due to climate change.
    b. Biodiversity loss and ecosystem collapse: The degradation of ecosystems can disrupt food supplies, increase the spread of diseases, and lead to loss of livelihoods, which in turn affect mortality rates.
    c. Critical changes to Earth systems: Climate tipping points, such as the melting of polar ice caps and shifts in ocean currents, can have far-reaching and potentially catastrophic impacts on human health and survival.
3. Geopolitical Risks.
    a. Conflict and instability: War and political instability can lead to direct mortality from violence, as well as indirect mortality through the destruction of healthcare infrastructure, displacement, and famine.

b. Misinformation and disinformation: The spread of false information, particularly regarding health and safety, can exacerbate public health crises and hinder effective responses to emergencies."

When I further queried it regarding which of these risks are most relevant to high-income countries, it came back with:

1. "Non-communicable diseases: heart disease and stroke, cancer, and diabetes.
2. Mental health and substance abuse.
3. Climate-related events, including extreme weather events, and air pollution.
4. Pandemics and infectious diseases: influenza and respiratory diseases, and antimicrobial resistance.
5. Lifestyle-related diseases: obesity and related complications, smoking, and alcohol use."

I could have then proceeded to ask, for each of these risks for further recent research to determine the likelihood that these risks will occur or become more severe, and what age groups they might affect most.

After a further inquiry regarding the best process used to project mortality improvement, it concluded that "Projecting future mortality improvements is a complex but manageable task involving careful analysis of historical data, health trends, and potential future changes in medical care and public health policies. By using robust statistical models and considering a range of scenarios, more accurate projections can be developed to inform public health planning and policymaking." Certainly, good advice for actuaries developing mortality improvement assumptions.

This suggests that, with further inquiries, AI can provide valuable input and assistance, listing the basic steps that an actuary could follow. However, the execution of each step should be handled or overseen by a well-trained and experienced actuary to complete an analysis or other project.

THE GUT CHECK

Sometimes a check for reasonableness is referred to as a gut check or intuition. Are the results (and intermediate findings) reasonable? In practice, these findings are often professionally peer-reviewed to determine if the relevant factors have been considered, appropriate processes followed, and are properly documented. These parts of model governance need to be incorporated and assessed in the use of any algorithmic approach. A review of the underlying data or information could focus on the data/information that the AI was trained on, if obtainable, with the results being thoroughly validated and checked for reasonableness.

I draw an analogy to predictive keyboarding – my word processor pre-shadows my typing in many cases to help me become more efficient, but its suggestions are wrong almost as often as they are right. At the same time, it also can be distracting. Similarly, if an actuary depends too much on a trained algorithm without proper supervision or reasonableness checking, the risk is that the findings, conclusions, or recommendations could be wrong or lead the actuary astray.

CONCLUSIONS

AI is a hugely important technical development, representing a potentially valuable tool that may be able to enhance the practice of an actuary. If mastered and properly directed, it can serve multiple purposes. Yet its algorithmic findings cannot replace the key functions of an actuary and may not even identify the initiating problem. In addition, it is unlikely to replace the identification and assessment of the quality of data/information source, the choice of the most relevant trajectory of analysis, and the assessment of the reasonableness of findings, especially in the development and use of scenarios. Just like other information development and modeling, it can enhance the efficiency of the processes followed. Nevertheless, the prospective nature of actuarial work demands checks and reviews throughout our processes. Without proper 'supervision,' over-reliance on an AI model can lead to failure and added risk.

POSTSCRIPT

After I finished this essay, I thought it appropriate to obtain a second opinion regarding the reliability of an AI model. To do so, I asked ChatGPT. Its conclusion was eerily similar to what I have discussed in this essay:

"Factors influencing AI model reliability:

1. Quality of training data: Diversity and representation. The training data must be diverse and representative of the real-world scenarios the model will encounter. Bias in the data can lead to unreliable and unfair outcomes, and quantity of data. More data generally leads to better performance, as the model can learn from a broader range of examples.
2. Model design and complexity: Algorithm choice: Different algorithms have different strengths and weaknesses. The choice of algorithm impacts the model's ability to generalize from the training data to new data. Overfitting/underfitting: A reliable model strikes a balance between overfitting (too closely matching the training data, reducing generalizability) and underfitting (failing to capture patterns in the training data).
3. Testing and validation: Cross-validation: Techniques like cross-validation help ensure the model performs well on unseen data by splitting the data into training and testing sets. Performance metrics: Metrics such as accuracy, precision, recall, F1 score, and area under the curve provide insights into the model's performance and reliability.
4. Robustness to changes: Adaptability: A reliable AI model can handle variations and changes in the input data over time. Stress testing: Models should be tested under a variety of conditions to ensure they are robust and reliable in different scenarios.
5. Transparency and explainability: Interpretability: Models that are easier to interpret and explain are generally more reliable because their decisions can be understood and validated by humans. Bias and fairness audits: Regular audits for bias and fairness help ensure the model's reliability and ethical integrity.

Reality in Practice

1. Healthcare: AI models in healthcare must be highly reliable as they can directly impact patient outcomes. Rigorous testing, validation, and regulatory compliance are critical. Example: AI models for diagnosing diseases from medical images have shown high accuracy in controlled studies, but real-world reliability requires extensive validation and oversight.
2. Finance: In finance, AI models are used for credit scoring, fraud detection, and trading. Reliability is crucial to avoid significant financial risks. Example: AI models for credit scoring must be continuously monitored to ensure they do not introduce bias against certain groups and remain accurate as economic conditions change.
3. Autonomous vehicles: Autonomous vehicle AI systems must reliably interpret and respond to a wide range of driving conditions to ensure safety. Example: Despite advancements, the reliability of autonomous driving systems is still a concern, with ongoing development needed to handle edge cases and improve overall safety.

Limitations and challenges

1. Data privacy: Ensuring data privacy while using large datasets for training can be challenging.
2. Bias and fairness: Addressing inherent biases in training data is an ongoing challenge, as biased models can lead to unfair and unreliable outcomes.

3. Dynamic environments: AI models may struggle with rapidly changing environments unless they are regularly updated with new data and retrained.

Conclusion

The reliability of an AI model is contingent upon careful design, thorough testing, and continuous monitoring. While AI models can achieve high levels of accuracy and reliability, it is crucial to address limitations such as bias, data quality, and adaptability to changing conditions to ensure they remain trustworthy and effective in real-world applications."

Now that I read what ChatGPT produced regarding AI models and eliminated the jargon, it looks similar to what I would say about actuarial modeling and practice, with the addition of the use of actuarial judgment and a profession-wide self-governance (in most cases) process.

* * * * *

Sam Gutterman, FSA, CERA, FCAS, MAAA, FCA, HonFIA is a consulting actuary. He can be reached at sam.gutterman1@gmail.com.

Plausible Risk of Structural Racism from the Widespread Use of AI

Rajeshwarie VS and Vivek Gupta

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

INTRODUCTION

Human intelligence[1] is the mental ability to learn and adapt to new situations, to understand and deal with abstract concepts, communication, reasoning, problem solving, and memory formation. In the past two decades, artificial intelligence (AI) has emerged as a powerful tool to fill in the gaps of human intelligence – speed and capacity. With its ability to process extremely large amounts of data in exceedingly short time spans, AI first began taking over repetitive and human energy/resource intensive tasks and evolved from automation to prediction and classification. The National Artificial Intelligence Act 2020 defines artificial intelligence as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Both definitions of human and artificial intelligence are not very different except for perhaps communication, problem-solving and adapting to new situations which are still unique human strengths. AI follows a human-like evolutionary path of learning and growing, first being taught/trained to look for patterns and then to start spotting them at a much larger scale and at phenomenal speed yet unmatched by the human brain. Humans learn their biases through prejudices, stereotypes and rhetoric that have and may still prevail in society, and not always from scientific data or study. To counteract this, States across the globe have enacted laws to protect vulnerable minorities from discrimination. These laws categorically prohibit use of inherent traits to deny/discriminate people in any offerings unless they prove differentiation is fair and based on actual scientific data. These laws are made with the social goal to have a fair and equitable society where protected classes are not left marginalized. When AI is fed with data that carries the imprint of human biases, there is a great chance that these biases will enter models, as advanced algorithms can find new proxies to confirm these data biases and create more unexplainable models that strengthen these biases.

The nature of insurance business is "discriminatory" in the sense that it differentiates policyholders based on information about them and classifies them into "good" and "bad" risks. Such risk classification in order to assess and charge premiums is inherent to insurance business and is justifiable. But to what extent can one discriminate is a key question. Laws exist that prevent using certain information as rating factors or as a basis for providing/rejecting coverage. Gender, race, pre-existing health conditions and genetic information cannot be used to deny coverage or charge different premiums.

---

[1] Human Intelligence - an overview | ScienceDirect Topics

While direct discrimination is prohibited by laws, there are still ways by which scrupulous parties can unknowingly discriminate against protected groups by creating surrogate variables or rules in their processes thus undermining the objective of legislation. Two prominent types in which it can occur are:

> Structural racism[2] is defined as social, economic, and cultural differences between different groups of people that have developed over time leaving patterns entrenched in society. So, although an insurance pricing model may not be built with race as one of the rating factors, the results of the pricing exercise may still be to the advantage of some groups and/or a disadvantage to some others.
>
> Disparate Impact[3]: Policies, processes, rules, or other systems that appear to be neutral, result in a disproportionate impact on the protected classes. An example would be a fraud score which uses zip code as one of the parameters as this may cause a higher cost to poor minority neighborhoods. Zip codes have no direct impact on fraud score of an individual and are most easily available information. But zip codes are proxies to income, ethnicity and race and therefore it is possible that using a seemingly neutral factor such as a zip code in an algorithm to assign a fraud score, may produce unfair outcomes to some socio-economic groups which could then result in higher borrowing costs or insurance premiums.

In the rest of this paper, we look at the possible biases that may creep in if an insurance company uses AI extensively in its customer-facing processes largely due to rating factors or socio-economic indicators that may unintentionally work as proxies for protected factors in AI algorithms. These algorithms are often proprietary and hence black boxes which makes it difficult to assess the prediction logic. We conclude with a review of nascent fields of regulation and proactive measures that are developing to use AI ethically.

AI IN INSURANCE: THE PROBLEM OF PROXIES AND BLACK BOXES

AI is revolutionizing insurance in a number of ways with the ability to process vast amounts of data and generate insights for pricing and risk evaluation, targeted marketing, claims handling and settlement, streamlining manual and repeatable tasks, thus speeding up the insurance process for all parties involved as well as reducing the cost of delivering services. By leveraging AI, insurance companies can create and customize insurance products as per the need of the insured and price it better in commiseration to risk. With data ingestion platforms, a lot of data entry tasks can be reduced for underwriters and claim handlers thus freeing up their time to focus more on value added services of risk evaluation and claim adjudication. As a logical next step, insurance companies might leverage AI to make decisions without human intervention or with very limited human oversight. There is a good possibility that it may have negative impacts on the way insurance companies treat the insured. A number of recent studies have pointed to such negative impacts and below are some examples.

A number of rating factors are used to underwrite and price risks. For auto insurance, traditionally, vehicle and driver related information have always been the rating factors, for example, vehicle make and model, vehicle and driver age, history of previous claims/accidents, mileage, and zip code. A number of states can use non-driving related factors such as gender, marital status, home ownership details, address or zip-code, education level, past insurance purchase history. While it may seem that some of these factors are obvious choices for determining driver behavior and the risk the drivers represent, some of these factors can act as proxies for others that are explicitly banned. A number of studies[4] point to drivers from predominantly Black/African American[5] U.S. neighborhoods being charged significantly higher premiums than those from white neighborhoods, whites who live in rented

[2] Definition of structural racism - NCI Dictionary of Cancer Terms - NCI
[3] Defining Fairness and Equity in AI-enabled Fraud Detection | Voyatek
[4] Study Points to Rate Bias in U.S. Auto Insurance Industry (investopedia.com)
[5] For brevity throughout the remainder of this essay, the author uses “Black” to represent “Black/African American.”

houses being charged lower premiums than Black homeowners, and Black drivers with better credit scores having higher premiums than white drivers with poorer credit scores. None of the algorithms that generate these rates have race as an input. Another more common example is exploiting the correlation between smoking and gender, while gender cannot be used as a rating factor, health and life insurance requires smoking status, and justifiably so!

While the above two examples illustrate the problem of proxies, there is also the problem of AI algorithms being black boxes, where it is not always obvious how the input data is being used to generate insights. An algorithm[6] used for predicting complex health care needs assigns a risk score to each individual based on past data of health care costs, and patients with risk scores above a threshold are automatically enrolled in the care program. The algorithm takes insurance claims data – age, sex, insurance type, diagnosis, procedures, and costs but explicitly excludes race. In predicting future healthcare costs based on the data, the algorithm performs consistently and predicts more or less equal costs for Blacks and whites irrespective of risk score. But the results showed that while referring people to support/care programs, a lower number of Blacks than whites at the same risk level were automatically selected. A study[7] on the results generated by the algorithm highlights that at the same level of health, Blacks tend to spend less on healthcare than whites and the kinds of expenses the two groups incurred were also different. Blacks spent more on emergency hospital visits while whites spent more on inpatient surgical procedures and specialist fees. While the algorithm was accurate in predicting future care costs per person, when it came to identifying care needs, using this predicted variable as an indicator of who needs care was failing to produce similar and equitable results for all groups.

DATA

AI is being used in claims settlement for initiating, assessing as well as approving claims, to quicken the process and offer customized responses to policyholders. Algorithms are widely used to file and screen claims and often involve processes such as checks to assess if forms have been filled correctly, images of damage are uploaded with required resolution, and screening basic information for possibility of fraud, thus saving many manhours. There have been studies that show that these practices have been discriminatory[8] with many instances of policyholders belonging to racial minorities being asked for additional information or documentation thus leading to a longer and more complicated settlement process. While no algorithms may have been built with the explicit aim to produce such delays, differences in levels of awareness, knowledge, digital literacy, access to uninterrupted internet, speed or resolution of devices that policyholders use may all impact the quality of the inputs they provide resulting in the longer delays, rejections, and requirement for additional validation. All these factors are known to vary with economic circumstances which vary widely between different ethnic groups and segments of society.

On the underwriting side, in order to classify and evaluate the risk represented by a prospective or existing policyholder, data is vital. In earlier times, this data was mostly available from the policyholders themselves and from the insurer's own experience. With advancement in AI, a large number of companies collect data and process it from various sources – shopping trends, credit scores, health parameters, travel, and food preferences are all accessible to insurers to deliver tailor-made products at accurate pricing. While we may believe that this data collected independently and through disjoint sources is unbiased and reflects lifestyle and behaviors of users as is, we need to understand that whatever insights it can provide would be limited by differences[9] in access, knowledge, awareness, ease/comfort with technology among different users or groups of users. It is well known and acknowledged that race and many socio-economic factors are correlated. For instance, in health care, poverty

[6] For minorities, biased AI algorithms can damage almost every part of life | SOAS
[7] Dissecting racial bias in an algorithm used to manage the health of populations | Science
[8] State Farm accused of covert racial discrimination in claims processing | WGLT
[9] The data divide | Ada Lovelace Institute

differences, lack of awareness to seek care, difference in access to transportation and health facilities, other competing demands like jobs or childcare, and doctor-patient relationships result in different levels in use of health care. Most of these factors would similarly impact adoption and optimal use of new technology and therefore any data collected through these will have an under-representation of some groups. While this data is used as such by AI/machine learning algorithms, these patterns are further amplified. Any prediction logic built with the majority, white-collared, educated and tech-savvy individual in mind would work well only for this segment of the population and may result in unfair results to any others that might not meet even one of these criteria.

REMEDIAL AND NEXT STEPS

The legal and regulatory landscapes have been constantly evolving to help counteract any inequality or discrimination that is perpetuated by technology. Beginning with the Fair Housing Act to counteract redlining, there are now many laws[10] such as Health Insurance Portability and Accountability Act of 1996, Genetic information non-discrimination act 2008, and Patient Protection and Affordable Care Act 2010 that apply to insurers to prevent any unjustified discrimination. More recently, states[11] in the U.S. are coming up with legislation specifically to avert unfair discrimination resulting from the extensive use of AI. These acts aim to ensure that an insurer's use of external consumer data sources and predictive models are augmented by a risk assessment and management framework that has to be documented and submitted to the regulators. Such frameworks require, on an ongoing basis, insurers to

- disclose external information sources and the algorithms or predictive modes they use,
- explain the manner in which both the data and the predictive models are used,
- assess where the use of these data sources and algorithms can cause unfair discrimination based on gender, race, color, ethnicity, disability, sexual orientation, and other prohibited factors.
- provide a reasonable time frame to remedy any such discriminatory impact of an algorithm.

Apart from the legal and regulatory aspects to prevent unfair discrimination, insurers can proactively work on the data and models to ensure the same. Companies could choose to work with the input data, the models themselves, the outputs or how they use these outputs in a manner to prevent any disparate impact.

- Data: A good understanding of the data that is used and how it has been collected is key to assessing the limitations and biases inherent in it and is a first step in eliminating biased outputs. The next step could be pre-processing input data, or post-processing outputs to mitigate the effects of biases that the data contains.[12]
- Models: It is important to understand what happens in AI models – their prediction or classification logic. Identifying the right relationship between the predicted variable and various predictors and testing this logic is necessary and is also a key area for regulators. Data insufficiencies or limitations can be corrected in the model training phase.
- Output: There is an emerging body of work on 'fairness criteria'[13] that can be used alongside AI models. Such criteria can determine if the output produced is fair to all racial/ethnic and other protected classes. Companies can adjust outputs to remove the effect of these biases.
- Using the output: In some situations, because of very limited availability of data and access to technology, some sections of consumers may be heavily underrepresented in model input data. In such cases, it is

---

[10] repository.law.umich.edu/cgi/viewcontent.cgi?article=1163&context=law_econ_current
[11] 2021a_169_signed.pdf (colorado.gov)
[12] How insurers can mitigate the discrimination risks posed by AI - UNSW BusinessThink
[13] Deloitte_Trustworthy AI_Fairness_Whitepaper_Dec2021.pdf

essential that companies review the applicability of model outputs to such groups. While it may work for some, it may not work for all!

CONCLUSION

In some areas such as criminal sentencing and facial recognition in public spaces, there is a demand to ban use of AI. In insurance, although it is still an area that people are suspicious about with the recent spate of lawsuits and literature around unfair discrimination, It is also widely acknowledged that there are benefits of AI in hastening up the various tasks along the insurance value chain that translate to savings in cost and resources. The alarm that is being raised is not merely resistance to change and should not be dismissed as voices against adoption to new technology. Companies and practitioners of AI need to acknowledge that the risks exist, and these concerns are legitimate as a first step to working on a solution to provide equitable and fair insurance offerings to all.

* * * * *

Rajeshwarie VS is a P&C actuary based in India. She can be reached at r.eshwari.894@gmail.com.

Vivek Gupta is an insurance consultant based in India. He can be reached to hutchie@gmail.com.

Catastrophic Risks of AI-Based Chatbots in Educational Systems

Ali M. Saghiri

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

ABSTRACT

AI-based chatbots are going to be used by a wide range of people, including students, teachers, and parents, for various purposes, and their impact on improving the productivity of teachers and students is undeniable. These systems will bring personalized learning for all students in different levels of education. However, it seems that rapid deployment without addressing the inherent challenges of AI-based systems may lead to several risks that have not been reported in the literature due to the fast rise of these systems. This essay investigates the potential catastrophic risks associated with the use of AI-based chatbots in educational settings, integrating insights from recent studies with a particular focus on privacy, security, ethical dilemmas, and technological dependency. Recognizing the emerging challenges, I propose a novel solution leveraging blockchain technology to enhance the security, transparency, and integrity of AI-based chatbots in educational environments. Recommendations for mitigating these risks are provided, emphasizing the unique context of educational institutions and the need for innovative approaches to safeguard student data and maintain educational quality.

INTRODUCTION

The integration of AI-based chatbots into educational systems offers promising enhancements in personalized learning and student engagement. However, these systems also introduce significant risks that could undermine educational integrity and security. With the increasing adoption of AI technologies in education, there is a growing concern about the privacy of student data, the fairness of algorithmic decision-making, and the potential for misinformation dissemination.

AI-based systems in educational environments present several challenges, particularly concerning privacy, security, and safety. Privacy concerns arise as these systems often handle sensitive information, including student performance data, personal details, and behavioral insights, which require robust protections against unauthorized access and breaches. Security is another significant issue, as the infrastructure supporting AI systems must be safeguarded against cyber threats that could compromise the integrity and availability of educational resources. Safety issues also surface, especially regarding the ethical use of AI in shaping educational content and interacting with students, where there's a risk of bias and misrepresentation. Ensuring these systems are transparent, accountable, and aligned with educational ethics is crucial to address these challenges effectively and maintain trust among students, educators, and parents.

In response to these challenges, I propose a novel solution that has not been extensively explored in this domain: leveraging blockchain technology to enhance the security, transparency, and integrity of AI-based chatbots in educational settings. Blockchain, a distributed ledger technology, offers unique features such as immutability, transparency, and decentralization, which can address the vulnerabilities associated with centralized systems.

By implementing blockchain-based solutions, educational institutions can create tamper-proof records of chatbot interactions, ensure the integrity of student data, and enhance trust in AI-driven educational tools. Furthermore, the use of consensus algorithms within blockchain networks can provide additional security and validation mechanisms, further mitigating the risks of data breaches and algorithmic biases. In this essay, I discuss the potential catastrophic risks posed by AI-based chatbots in educational systems and propose innovative strategies leveraging blockchain technology to address these challenges. Through collaboration between educators, technologists, and policymakers, I aim to promote the adoption of blockchain-based solutions to safeguard student data and uphold the quality of education in the digital age. The next two sections are dedicated to the details of risks and potential solutions to elaborate the importance of innovative solutions in this field.

DISCUSSION OF CATASTROPHIC RISKS

One of the primary concerns is the handling of sensitive student data, which raises serious privacy and security issues. AI systems, if not adequately protected, are susceptible to breaches that could lead to significant violations of privacy laws, risking both student trust and institutional credibility. Furthermore, ethical considerations must be meticulously managed; biases inherent in algorithms could skew evaluations and recommendations, potentially leading to unfair academic outcomes. Such biases could perpetuate or even exacerbate existing educational inequalities, highlighting the need for continuous oversight and correction of these intelligent systems.

Another significant risk is the potential over-reliance on chatbots, which, while efficient, might diminish critical thinking skills among students. If these tools are overly relied upon for educational interaction, students may miss out on the benefits of direct human engagement, which fosters deeper understanding and critical analysis. Moreover, the risk of misinformation is substantial; inaccuracies propagated by AI systems can mislead students, distort educational content, and impair learning outcomes. These risks underscore the necessity for a balanced integration of AI technologies with traditional teaching methods, ensuring that AI complements rather than replaces human interaction in education.

POTENTIAL SOLUTIONS AND MITIGATIONS

To address mentioned risks, educational institutions should implement comprehensive cybersecurity measures, develop ethical guidelines for AI use, and maintain an active oversight committee to monitor AI integration. Educator training on AI capabilities and limitations is also recommended. Moreover, collaboration between educators, technologists, and policymakers is essential to develop effective solutions. Incorporating blockchain technology can enhance the security and transparency of AI-based chatbots in educational systems. By leveraging blockchain, educational institutions can create immutable records of chatbot interactions, ensuring data integrity and reducing the risk of tampering or unauthorized access. Additionally, consensus algorithms, such as proof of authority or proof of stake, can be implemented to validate transactions and maintain the integrity of the blockchain network.

CONCLUSION

In conclusion, the integration of AI-based chatbots in educational systems presents both opportunities and risks. I proposed leveraging blockchain technology to enhance the security and transparency of these systems. By creating tamper-proof records of interactions and implementing consensus algorithms, we can mitigate risks and safeguard student data. Collaboration and further research are essential to ensure the responsible use of AI technologies in education. With innovative solutions like blockchain, we can promote a secure and transparent educational environment, supporting positive learning experiences for students.

* * * * *

Ali M Saghiri is an assistant professor at William Paterson university. He has published more than 50 scientific papers in international conferences and journals, two books, and five book chapters. His research interests include the Internet of Things, blockchain, and artificial intelligence. He can be reached at saghiria@wpunj.edu.

REFERENCES

1. Ali M Saghiri, et al., "A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions." *Applied Sciences*, vol. 12, no.8, 2022, p. 4054, DOI: 10.3390/app12084054.

2. Holmes, Wayne, et al. "Ethics in AIED: Historical perspective and future directions." *International Journal of Artificial Intelligence in Education* 29, no. 3 (2019): 315-334.

3. Shum, Simon Buckingham, and Rose Luckin. "Learning analytics and AI: Politics, pedagogy, and practices." *British Journal of Educational Technology* 50, no. 6 (2019): 2795-2808.

4. He, Jieun, et al. "The potential and risks of AI-driven educational chatbots in special education." *Journal of Computer Assisted Learning* 37, no. 1 (2021): 82-95.

5. Roll, Ido, and Ryan S. Baker. "The emerging role of educational data sci- ence in education and learning science." *The Cambridge Handbook of the Learning Sciences,* (2020): 524-547. 3

6. Drachsler, Hendrik, and Wolfgang Greller. "Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics." *Proceedings of the Sixth International Conference on Learning Analytics Knowledge*,2016.

7. Bittencourt, Igor I., et al. "The educational affordances of multimodal serious games and intelligent tutoring systems: A systematic review and a synthesis of the literature." *Journal of Computer Assisted Learning* 34, no. 6 (2018): 614-626.

8. S. S. Gill et al., "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots." *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 19–23, 2024.

9. M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A SWOT analysis of ChatGPT: Implications for educational practice and research." *Innovations in Education and Teaching International*, pp. 1–15, 2023.

10. S. Gökçearslan, C. Tosun, and Z. G. Erdemir, "Benefits, challenges, and methods of artificial intelligence (AI) chatbots in education: A systematic literature review." *International Journal of Technology in Education*, vol. 7, no. 1, pp. 19–39, 2024.

11. E. Guustaaf, U. Rahardja, Q. Aini, H. W. Maharani, and N. A. Santoso, "Blockchain-based education project." *Aptisi Transactions on Management (ATM)*, vol. 5, no. 1, pp. 46–61, 2021.

Artificial Intelligence Discrimination: Cause, Damage and Mitigation

Kailan Shang, FSA, CFA, PRM, SCJP

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

As a simulation of human intelligence, artificial intelligence (AI) has been used to perform tasks with limited human intervention. The recent breakthrough of large language models (LLMs) allows us to use AI with little instruction. AI can act as a helpful and creative adviser and significantly improve our productivity. At the same time, the easy access to AI may lead to unexpected and undesired outcomes when there is a lack of controls and experience.

One of the benefits of using AI is to make unbiased decisions. Unlike humans, AI is not driven by emotions and adopts rational approaches. For example, an AI program can be built to use skills and qualifications to screen job applicants. Without demographic information, the recruiting process is expected to be more inclusive. Ironically, as studied in Chen (2023), biases in terms of gender, race, skin color, and personality were present in AI powered recruitment processes. Other biases are not uncommon in AI, and they need to be understood and addressed to avoid wide and adverse impact on our societies.

AI BIASES: EXAMPLES AND IMPACT

Biases in the AI programs may be observed in different areas such as the training data, the algorithm used for predicting, and the predictions themselves. It is no surprise that our data contains biases as it is measured and collected by humans who are subject to different biases. Algorithms can sometimes reinforce the biases in the data to a greater extent, ignoring new patterns in new data. When biases can be easily observable in prediction results, lack of robust validation becomes obvious.

Table 1
EXAMPLES OF AI BIASES

Area	Description	Example
Training Data	Using biased data to train AI models, the biases are likely to be kept in the AI algorithm. • **Selection bias**: the data is not representative of the population under study due to incomplete data, biased sampling, and so on. • **Measurement bias**: the data collected differs systematically from the reality due to a measurement issue. • **Prejudice bias**: the data includes existing human stereotypes and assumptions	U.S. hospitals used an algorithm to predict the need of extra medical care. Historical health care spendings were used as a measurement of the needs. This inappropriate measurement caused issues to underestimate the medical needs of black patients.[1]
Algorithm	Although the training data may not contain demographic information as the source of biases, the model may learn from highly correlated variables and unintentionally discriminate against a certain group.	Amazon's AI-enabled hiring algorithm favored male applicants based on words "executed" or "captured" commonly used by men, and penalized resumes with the word "women's". The program discontinued after the findings. [2]
Prediction	The biases in predicted results are not always obvious. But when it is obvious, the impact is usually devastating to the business and to the technology.	Google's Gemini AI image generator produced images of historical figures in wrong and often darker skins. This led to the pause of this AI service. [3]

Example Sources:
[1] https://www.science.org/doi/10.1126/science.aax2342.
[2] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/.
[3] https://blog.google/products/gemini/gemini-image-generation-issue/.

When biased data is used for training AI models, it is challenging to predict the rare cases. Using the selection bias as an example, the data records that belong to the rare classes may be insufficient using standard processes. The training algorithm may be overwhelmed by the common cases and provide little insights on the rare cases. In addition, statistical measures may indicate a high level of prediction accuracy although rare cases are not predicted at all. For classification, precision, recall and the F-measure are popular measures based on the confusion matrix, as shown in Table 2.

Table 2
CONFUSION MATRIX ILLUSTRATION

	Predicted: True	Predicted: False
Actual: True	True Positive	False Negative
Actual: False	False Positive	True Negative

Precision measures the Type I error [1] and recall measures the Type II error. F-measure (or F-score) is the harmonic average of precision and recall and may be used as a high-level measure to rank the performance of different models.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall\ (True\ Positive\ Rate) = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

If we use the common class to calculate these measures, the prediction results may look promising, as shown in Table 3.

Table 3
CONFUSION MATRIX FOR THE COMMON CLASS

	Predicted: True	Predicted: False
Actual: True	95	0
Actual: False	5	0

Precision is 95%. Recall is 100%. F-measure is rounded to 97.4%. However, using the rare class, the same measures give us an opposite picture, as shown in Table 4.

Table 4
CONFUSION MATRIX FOR THE RARE CLASS

	Predicted: True	Predicted: False
Actual: True	0	5
Actual: False	0	95

Precision, recall and the F-measure become 0. It is clear that the AI model did a terrible job predicting the rare class mainly due to data imbalance in this example.

AI bias is not uncommon in the real applications. Even for companies that stay on top of the AI technology, it is possible that the AI applications may end up reinforcing immoral stereotypes and perpetuating inequities in our world. The impact can be quick, material and widespread, given the high efficiency of adopting AI in various areas in our societies. Certain groups may be given reduced opportunities in the economy. This can lead to a higher degree of economic inequality and social unrest. The trust in the technology may also be lost. With bad reputation and legal consequences, this promising technology may be ditched by business and society. Technological regression

[1] Recall from classical statistics, a Type I error is a false positive where you reject a true hypothesis. A Type II error is a false negative and occurs when you fail to reject a false hypothesis.

happened in human history. The collapse of the Roman Empire led to a decline in engineering projects, writing, and urbanization. Lower productivity is not an implausible scenario if there is a regression of AI technology.

AI BIAS IN INSURANCE

Many areas in the insurance industry have AI applications to facilitate automated decision making, including insurance pricing, underwriting, claim processing, and risk management. Bias of AI in insurance applications is no different from others.

- Insurance pricing. AI models may generate higher premium rates for certain demographic groups unintentionally. For example, insurance pricing models may use postal codes or location as a pricing factor. However, historical data may suggest a strong relationship between location and ethnic group. Demographic groups in poor communities are less likely to get affordable coverage. This means that ethnicity is indirectly used as a pricing factor as well. AI models may generate an unfair higher premium rate, even if the applicant's individual factors suggest otherwise due to data linking these areas to poorer health outcomes
- Underwriting. Similar to insurance pricing, AI bias may lead to unfair high-risk rating of minority groups even though individual data suggests otherwise. This can lead to limited access to insurance products and may affect the economic and social activities of the affected groups.
- Insurance claim. Claim processes may be different due to AI bias with certain groups having longer settlement time as more scrutiny was suggested to certain groups based on limited data that cannot differentiate further at individual claim levels.
- Risk management. AI-enabled fraud detection algorithms may target certain groups disproportionately suggested by training data but in reality, lead to wrong flagging of normal transactions.

Like other industries, AI bias in insurance applications may be caused by using datasets that are not representative of the entire insured group. Past practices may cause insurance data embedded with human biases. Without proper treatment, they will lead to a biased algorithm. When designing AI models, human bias may affect their fairness unintentionally.

MITIGATION STRATEGIES

When the training dataset contains selection bias, there are ways to address them in AI model training.

- Balancing the dataset. If your data volume is big enough, you may consider removing some data records that belong to the common class(es) to make is more balanced. On the other hand, if no data can be sacrificed, you can use oversampling to increase the number of data records belonging to the rare class(es). The new data records can be created by adding small noises to existing data records. Well-established algorithms such as the synthetic minority over-sampling technique (SMOTE) can be used to generate synthetic samples. The algorithm chooses a few similar data records like in clustering analysis and adjusts the explanatory variables by random amount limited to the difference to similar records.
- Adjusting the error function. The error function is the objective function to minimize in the model training process. It can be adjusted to penalize false negative cases with a heavier weight.
- Collecting more data records belonging to the rare class(es) if possible. This may be better achieved with efforts from the entire insurance industry.
- Categorizing the common class into subclasses to achieve balance through more classes, if possible.
- Using measures such as Receiver Operating Characteristic (ROC) that consider the prediction results of all classes, rather than a chosen class. The ROC curve helps understand the trade-off between the true positive rate and the false positive rate by varying the threshold that is used to determine whether a prediction is positive or negative.

Although from the technical perspective, methods are available to fight against AI bias, during the implementation of AI applications, insufficient efforts may be spent on AI bias due various reasons. It is therefore important to adopt good practices to ensure that AI bias is managed actively. Figure 1 lists some practices to mitigate the risk of AI bias.

Figure 1
AI PRACTICES TO MITIGATE RISK OF BIASED DECISION

Diverse Data	•Ensure training data is representive of the population under study •Collect data from multiple sources. Identify and remove biases in data
Fair Algorithm	•Set up formal processes aimed to assess fairness, equity and inclusion •Design algorithms that explicitly minimize bias
Transparency	•It is difficult to build trust in unbiasness with a black box. •Transparency helps ensure that unbiased data and alogorithms are used
Compliance	•Set up governance, and roles and responsibilities around AI bias •Follow relevant industry regulations and legal requirements
Expertise	•Build expertise on AI and its associated bias •Be familiar with the capabilities and limitations of AI and be honest about them
Risk Management	•Set up internal controls and regular monitoring •Allows human oversight in AI-enabled decision-making processes

In addition, domain knowledge and diversity in talents can provide us with different and relevant perspectives to fight against AI bias.

CONCLUSION

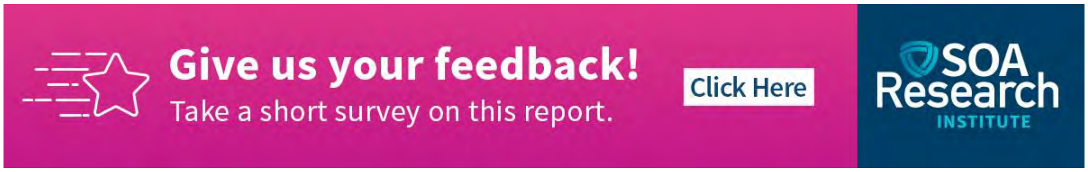
Given the remarkable advancements in AI, it is now easily accessible to the public with little prerequisites. At the same time, it also brings challenges to address the accompanying risks. In particular, AI bias can lead to unfairness in the decision-making process due to biased data and not well-designed algorithms. Bias was observed in AI-enabled health care, recruitment, and image generation systems. In the insurance industry, AI bias can affect fairness in underwriting, pricing, claim processing, and risk management. The impact of AI bias can be quick, material and widespread, given the high efficiency of adopting AI in various areas in our societies. Certain groups may be given reduced opportunities in the economy. This can lead to a higher degree of economic inequality and social unrest. The trust in the technology may also be lost. Lower productivity is not an implausible scenario if there is a regression of AI technology. The risk of systematic AI bias needs to be and can be addressed using technical approaches and sound AI practices.

* * * * *

Kailan Shang is a director at Aon PathWise. He can be reached at klshang81@gmail.com.

REFERENCES

Chen, Zhisheng, "Ethics and discrimination in artificial intelligence-enabled recruitment practices." *Humanities and Social Sciences Communications* 10, 567 (2023). https://www.nature.com/articles/s41599-023-02079-x.

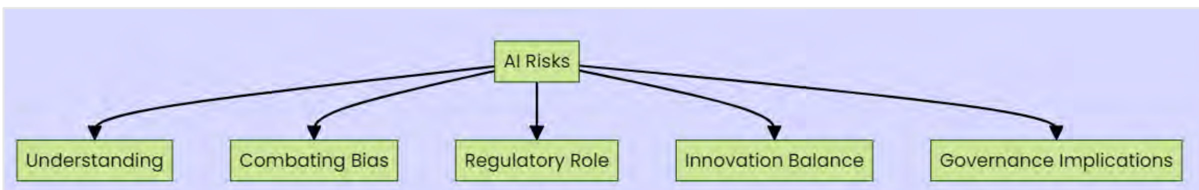

Give us your feedback!
Take a short survey on this report.
Click Here
SOA Research
INSTITUTE

Mitigate Biased Decision-Making in AI Algorithms

Jiaming Zuo, FSA, CERA, FCAA, FASHK

Any views and ideas expressed in the essays are the author's alone and may not reflect the views and ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, nor the author's employer.

The widespread use of artificial intelligence (AI) in the insurance sector brings several risks and challenges that companies need to navigate to ensure the responsible and ethical deployment of AI technologies. Here are some key risks emerging from the use of AI in the insurance industry.



Addressing algorithmic bias in AI systems requires a combination of technical approaches, ethical considerations, and regulatory oversight. Techniques such as bias detection, data preprocessing, fairness constraints, and explainable AI can help mitigate bias and promote more equitable and transparent decision-making in AI systems. By understanding how bias manifests and taking proactive steps to address it, developers and users can work towards building fairer and more inclusive AI technologies.

SEVERAL FACTORS CAN CONTRIBUTE TO ALGORITHMIC BIAS IN AI SYSTEMS

Several factors can contribute to the emergence of algorithmic bias in AI systems. These factors often intersect throughout the AI development process and can influence the presence and extent of bias in the resulting algorithms. Here are some key factors that contribute to the emergence of algorithmic bias:

Biased Training Data

The most common source of algorithmic bias is biased training data. Historical data often reflects societal biases, stereotypes, or systemic inequalities, which can be unintentionally encoded into AI systems during the training phase.

For example, an insurance company uses an AI algorithm to determine car insurance premiums for policyholders based on historical claims data. The training data predominantly consists of claims data from urban areas, leading to overrepresentation of claims from city drivers. The dataset lacks sufficient data from rural areas and under-represents low-income individuals.

Data Selection Bias

Data selection bias occurs when certain groups or perspectives are underrepresented or overrepresented in the training data, leading to skewed or incomplete datasets that do not accurately reflect the full range of real-world scenarios. The Imagin insurance company uses an AI algorithm to assess health insurance risk profiles based on historical claims data. The training data primarily consists of claims data from individuals who have regularly visited healthcare facilities and have a higher documented medical history. The dataset lacks representation of healthy individuals or those who may have had minimal healthcare needs.

Data Labeling Bias

Biases can also be introduced during the data labeling process, where human annotators may unknowingly inject their biases into the training data through subjective or culturally influenced labeling decisions. An insurance company uses an AI algorithm to assess risk profiles for home insurance policies. The algorithm relies on labeled data to identify risk factors associated with properties. The data labeling process involves human annotators who unintentionally introduce biases in determining property risks based on subjective judgments or assumptions.

Algorithm Design Choices

Algorithmic bias can be unintentionally introduced through design choices such as feature selection, model complexity, hyperparameter tuning, or optimization strategies. Biased assumptions embedded in the algorithm design can lead to biased outcomes.

Feedback Loop Effects

AI systems that interact with users and learn from feedback data can develop feedback loop bias. If the feedback data is biased, the system may reinforce or amplify existing biases over time, leading to discriminatory outcomes.

Contextual Biases

The context in which AI systems are deployed can also contribute to algorithmic bias. Biases may emerge from specific use cases, application domains, cultural norms, or social structures that influence the data collection, algorithm design, or decision-making processes.

Human Involvement

Humans involved in the AI development lifecycle, including data scientists, engineers, and designers, can introduce biases consciously or unconsciously. Their subjective judgment, prior beliefs, assumptions, or cultural influences can shape the AI system's behavior.

Lack of Diversity

Lack of diversity in AI development teams or insufficient representation of diverse perspectives and voices can contribute to the perpetuation of biases in AI systems. Diverse teams can bring different viewpoints and experiences to identify and address bias effectively.

Addressing algorithmic bias requires a holistic approach that involves careful data curation, transparency in algorithmic decision-making, diversity in AI teams, ongoing monitoring for bias, and the integration of fairness considerations throughout the AI development lifecycle. By understanding and mitigating the factors that contribute to bias, developers and practitioners can work towards creating AI systems that are more equitable, accountable, and inclusive.

THE CONSEQUENCES OF ALGORITHMIC BIAS ON ACTUARIAL ANALYSIS

In actuarial analysis, AI algorithms trained on biased data may lead to discriminatory outcomes in insurance underwriting and pricing practices. One example of this is the use of historical claims data that may contain inherent biases related to factors like race, gender, or socioeconomic status. If AI algorithms are trained on this biased data, they may inadvertently perpetuate these biases in insurance risk assessments and pricing decisions, leading to discriminatory outcomes for certain demographic groups.

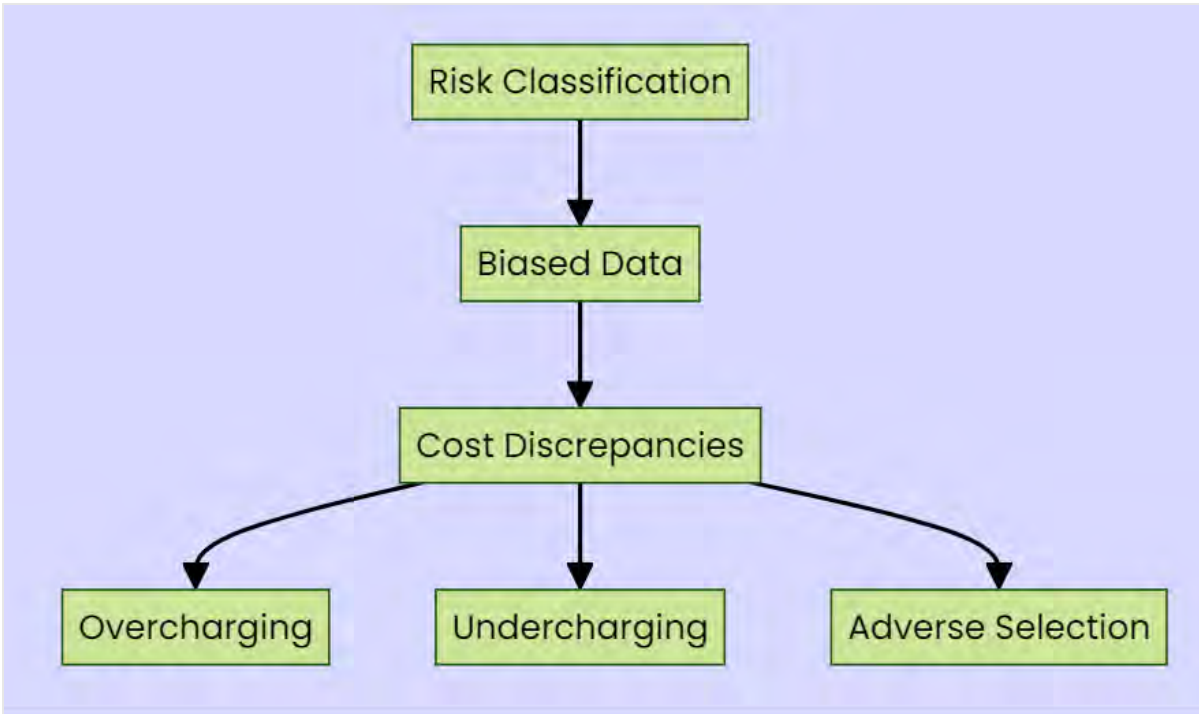
As noted above, actuarial services are data-driven; data bias, left unaddressed, can lead to incorrect conclusions, unwanted consequences, wrong policy decisions, or inadequate system performance. This section provides a few examples of actuarial services that can be impacted by data bias.

RISK CLASSIFICATION

Risk classification is the process of evaluating and estimating the future costs related to transferring risk. Biased data can introduce discrepancies between the actual future costs and the actuary's projections, potentially leading to overcharging or undercharging and adverse selection. Availability bias and historical bias are two significant factors that can impact actuarial decisions.

Using a machine learning model to predict life insurance policyholder mortality rates, an insurer inadvertently incorporates biased data that skews towards affluent applicants. As a result, the algorithm may underestimate risks for certain demographic groups, leading to improper risk assessments and potentially unsustainable pricing strategies. An additional instance of bias is historical bias, where differences in homeownership by race are overlooked in a personal auto rating plan. This omission can lead actuaries to base results on this bias rather than the genuine driver of future loss performance.

The Imagin insurance company uses an AI algorithm to assess risk profiles and determine premiums for auto insurance. If the algorithm is trained on biased data that correlates accidents with a specific demographic group rather than driving behavior, it may unfairly penalize individuals in that group with higher premiums, leading to discrimination and perpetuating unfair practices.



By paying more attention to possible historical influences in the data, the actuary can focus on the true drivers of future expected costs such as experience and driving record.

EXPERIENCE STUDIES

Experience studies are instrumental for life insurers in establishing accurate assumptions for life and annuity policy premiums. By aggregating mortality data from life companies to generate industry mortality tables, insurers inform their premium calculations. This process is mirrored in lapse assumptions, aiding insurers in comprehending policyholder decrements beyond mortality factors. A comprehensive understanding of unbiased historical data is paramount to mitigate the risks of poor assumption development and underpricing life insurance policies.

An AI system is used to automate claims processing for health insurance. If the algorithm is biased to associate certain medical conditions with higher costs or lower validity, it may systematically deny or delay legitimate claims from individuals with those conditions, resulting in unfair treatment and negative financial impacts for affected policyholders.

One more example, imagine an insurance company using an AI algorithm to analyze historical claims data for setting insurance premiums. The AI algorithm is trained on past claims data that inadvertently reflects biases against certain demographic groups, such as age or gender. Due to this biased training data, the algorithm may learn patterns that unfairly penalize older policyholders by overestimating their risk levels compared to younger policyholders.

RESERVING

The reserving actuary's core task is to estimate future claims and expenses by analyzing claims data and other experiences to establish crucial assumptions. In the domain of property and casualty insurance, the claims department is responsible for managing loss payments and reserves for future loss and expense payments. However, reserve analyses may encounter aggregation bias, particularly when attempting to generalize development patterns across diverse data subsets such as long-tailed liability and short-tailed property data.

MODELING

Actuaries employ modeling and advanced analytical techniques to refine decision-making in insurance operations. Omitted variable bias poses a substantial threat to risk classification models, as the exclusion of critical variables can introduce spurious correlations or signal loss, resulting in less comprehensive models and potentially skewed coefficient estimates. Similarly, confirmation bias can hinder predictive modeling efforts, prompting actuaries to manipulate models until they align with preconceived expectations, influencing assumption selection in reserving practices.

For instance, an insurance company implements AI-powered dynamic pricing for home insurance based on property features and location. If the algorithm incorporates biased assumptions about neighborhood characteristics or housing types, it may inadvertently undervalue or overvalue certain properties, leading to price disparities that disproportionately affect specific groups of policyholders.

REDUCING AND MITIGATING ALGORITHMIC BIAS IN AI SYSTEMS

Reducing and mitigating algorithmic bias in AI systems for actuarial analysis is essential to ensure fair and accurate decision-making processes. Here are some approaches and techniques that can help address algorithmic bias in actuarial analysis:

Diverse and Representative Data

To enhance fairness in model predictions, it is imperative to utilize diverse, inclusive, and representative training data for AI models, minimizing biases and ensuring equitable outcomes

Data Preprocessing

To fix biases in training data, use data preprocessing techniques like de-biasing, cleaning, feature tweaking, and balancing to ensure fair model training.

Fairness Constraints

Ensure fairness in AI algorithms by adding constraints to prevent discrimination during training, stopping biased patterns, and ensuring fair outcomes for everyone.

Explainable AI (XAI)

Use explainable AI (XAI) techniques to make AI models clearer and easier to understand. XAI shows how algorithms make decisions, helping to find and fix biased patterns.

Bias Audits

Conduct regular bias audits to assess and identify biases in AI models. Evaluate model performance across different demographic groups and identify disparities that may indicate algorithmic bias. Adjust the model parameters as needed to mitigate bias.

Human Oversight

Incorporate human oversight into AI systems to review and interpret model decisions, especially in sensitive or high-stakes applications such as actuarial analysis. Human intervention is critical for ensuring ethical and transparent AI-driven decision-making.

Sensitive Feature Removal

Remove or de-emphasize sensitive attributes (such as race, gender, or age) from the input data used for training AI models to prevent unwanted biases from influencing model predictions.

Regular Monitoring and Evaluation

Continuously monitor and evaluate AI systems for bias post-deployment. Implement feedback mechanisms, conduct bias testing, and solicit diverse perspectives to ensure fair and equitable outcomes over time.

Diverse Teams and Stakeholder Engagement

Encourage diversity and inclusion in AI teams to bring different perspectives to AI systems. Involve stakeholders, including those affected, to get feedback and ensure AI systems are ethical and fair.

By implementing these approaches and techniques, insurance companies and actuarial teams can work towards reducing algorithmic bias in AI systems used for actuarial analysis. Prioritizing fairness, transparency, and inclusivity in AI development processes can help build more reliable, ethical, and equitable AI-driven decision-making in the insurance industry.

* * * * *

Jiaming Zuo is a Senior Partner for EverBright Actuarial Consulting Limited. EverBright integrated AI into their digital platform for customizing and managing group health and insurance policies for clients. She can be reached at jzuo@ebactuary.com.

About The Society of Actuaries Research Institute

Serving as the research arm of the Society of Actuaries (SOA), the SOA Research Institute provides objective, data-driven research bringing together tried and true practices and future-focused approaches to address societal challenges and your business needs. The Institute provides trusted knowledge, extensive experience and new technologies to help effectively identify, predict and manage risks.

Representing the thousands of actuaries who help conduct critical research, the SOA Research Institute provides clarity and solutions on risks and societal challenges. The Institute connects actuaries, academics, employers, the insurance industry, regulators, research partners, foundations and research institutions, sponsors and non-governmental organizations, building an effective network which provides support, knowledge and expertise regarding the management of risk to benefit the industry and the public.

Managed by experienced actuaries and research experts from a broad range of industries, the SOA Research Institute creates, funds, develops and distributes research to elevate actuaries as leaders in measuring and managing risk. These efforts include studies, essay collections, webcasts, research papers, survey reports, and original research on topics impacting society.

Harnessing its peer-reviewed research, leading-edge technologies, new data tools and innovative practices, the Institute seeks to understand the underlying causes of risk and the possible outcomes. The Institute develops objective research spanning a variety of topics with its strategic research programs: aging and retirement; actuarial innovation and technology; mortality and longevity; diversity, equity and inclusion; health care cost trends; and catastrophe and climate risk. The Institute has a large volume of topical research available, including an expanding collection of international and market-specific research, experience studies, models and timely research.

Society of Actuaries Research Institute
8770 W Bryn Mawr Ave, Suite 1000
Chicago, IL 60631
www.SOA.org