Mphasis
The Next Applied

Explainable AI for Mitigating Biases in Judgment Systems

Whitepaper by Dr. Ashwani Singh, Manager – Applied AI NEXT Labs

Contents

1.

Human judgment systems augmented through AI

Across the world, numerous systems depend on human decisions to make judgments that subsequently affect the lives of other people. Loan Officers make judgments regarding loan recommendations, HR Managers judge which candidates to call for interviews, insurance providers evaluate whether a claim should be investigated further, Police Officers identify whether to search a vehicle or not and so on.

Recent developments have seen many of these judgments being replaced or supported by AI/ML systems, where a machine learning model makes the decision – partly or wholly. Banks and Financial Institutions are increasingly relying on AI-based solutions to improve the complicated task of predicting which individuals and entities represent the best risk to reward tradeoffs. Insurance industries use intelligent systems to identify potential fraudulent policy claims and flag the ones which are unlikely to be fraudulent to reduce the burden on investigators. HR Departments have started utilizing AI-based predictors to limit the applicant pool for positions. Law enforcement organizations have also been using them to estimate the probability of convicts committing crimes upon parole.

2.

Human judgments are prone to systematic biases that AI systems can perpetuate

As machine learning components take over parts or whole of decision making from humans, there have also been calls to ascertain if the algorithms being utilized are biased or unfair in any way. Unbiased and fair algorithms are expected to go a long way in earning the trust of individuals who are recipients of AI-based judgments and thereby increasing acceptance of these systems in the future. Unfortunately, errors in human judgments themselves are the major cause of bias in AI algorithms that are created to augment or replace these judgments.

Humans routinely make decisions that not just fail, but fail predictably. According to Daniel Kahneman[1] (Nobel prize for Economics 2002), the human brain employs two systems of thinking. System 1 is automatic, intuitive and fast paced with multiple mental modules working in parallel to produce a solution, for instance judging whether someone is trustworthy, within seconds of taking the first look at them. System 2 is deliberative, rational, slow paced and sequential, for instance, medical professionals narrowing down on diagnosis through the systematic elimination of alternative diagnostics. The quick and intuitive System 1 allows humans to handle various contexts and situations at a fast pace, making them prone to many biases.

These biases can produce judgments that have an inadvertent negative impact on certain groups. Biased thinking can result in some racial groups paying higher interest rates for loans and mortgages, being denied loans or insurance. Biases may also lead to unconscious discrimination against a particular sex for hiring decisions. Similarly, judges may disproportionately identify certain groups as representing a higher risk of criminal behavior upon release, which diminishes their chances of getting parole.

Machine learning systems depend upon historical data comprising input variables and consequent outcomes to train the models to augment or replace human judgments. If the models are trained on biased historical data and decisions, they would continue to perpetuate the same biases inadvertently in the future. This is especially true if the designers of the AI models share the same characteristics as the decision makers. So, if the data scientists are predominantly white males, they may neither be cognizant of nor identify biases against women and non-whites and are likely to find nothing untoward regarding patterns existing in historical data. This in turn could lead AI systems to predict outcomes that adversely affect a race or sex, leaving organizations vulnerable to regulatory and legal action, especially if the model is taking critical decisions that have a far-reaching impact on individuals. Thus, organizations that are looking at replacing or augmenting human systems by AI, need to understand what kind of biases may be perpetuated through AI systems and take steps to mitigate them.

3.

Explainable AI for bias-free AI systems

The recent advent of explainable AI components, which help users to understand the data and visualize how models are making predictions based on the data, could be useful in designing bias-free systems. Explainable AI or XAI has come up in the past few years as a response to calls for greater accountability of AI systems replacing or supplementing human judgments. Technology leaders have expressed the need to better understand how these decisions are made before they can be trusted to take over from humans. Indeed, almost[2] two-thirds of senior executives consider the ability to trace reasoning paths of AI predictions to be important. Moreover, regulatory agencies like the European Commission have also taken cognizance of the need for understanding how models behave and provided guidelines[3] for trustworthy AI solutions that depend heavily on model interpretability. XAI solutions address these concerns by incorporating components to clarify the reasons behind a machine's predictions, thereby reducing the black box nature of such models. Thus, XAI is a priority area for organizations that are looking at AI systems for replacing or supplementing humans in their processes.

XAI can also help in the development of bias-free AI systems through the incorporation of specific explanation strategies into interpretation modules of AI solutions to mitigate some common cognitive biases. This would thereby lead to better decision making, which in turn should reduce adverse outcomes from human judgment biases and improve outcomes for end users. Moreover, assurance of fairness concerns being met should reduce apprehensions about the adoption of AI systems and increase trust in machine-based judgments.

4.

Identifying and mitigating common biases in judgments through explainable AI

The table below identifies some of the most common biases and errors, and provides specific explanation strategies that can be incorporated into AI-augmented systems to identify and mitigate these biases.

Bias	Bias identification through explainable AI	Bias mitigation through explainable AI
Representativeness Bias: The user mistakenly perceives the probability of an outcome to be high due to the perceived similarity between two cases, leading to a wrong classification/decision. Resorting to easy stereotypes may be considered a form of representativeness bias where the user or decision maker pays a lot of attention to one feature (race/ethnicity/sex/age) and not enough to other relevant features while deciding. For instance, if loan disbursers reject low risk Black/Hispanic loan seekers with credit histories similar to loan winning white candidates, they may be judging individual applicants on the basis of negative impressions of the general population.	Counterfactual checks, which involve small changes in input values to check if the predicted values change can be employed in this case. If there's suspicion of bias in historical data, only the feature in focus (sex/race) of the case can be changed to see if the decision/outcome changes. Hence if changing Black to White, or Hispanic to White in case of applicants changes the predicted outcome, the judge may have been prone to this bias.	This bias can be mitigated by showing contrasts between the current case and an average case via a similarity distance measure. Hence, the individual's difference from the average on relevant criteria may be highlighted for the judge before making a judgment. This way, a judge's impressions based on what the class (Black/ Hispanics) represents would be adjusted through information on the individual that would otherwise not be considered.
Availability Heuristic: This occurs when the user overestimates the probability of occurrence of an event due to several similar instances happening around the user. For instance, if there are many stories of individuals committing insurance fraud in the news, or a major story involving insurance fraud in a particular city is given large coverage in the news, insurers could disproportionately identify claims from that city as potentially fraudulent and worthy of investigation.	Explainable AI providing relative weights of features considered while making the decision can be used to compare feature weight with news stories, to check if the relative weight of the city in focus changes with the news stories.	Explanations containing prior probabilities of occurrences should give users a more accurate picture of reality and mitigate availability bias.

In-group Bias: Here, the user favors members of a group the user belongs to (in-group), and disfavors members of other groups (out-group), resulting in better evaluation, higher allocation of resources for the in-group as compared to the out-group. For instance, if the employee making the hiring decision for a coding job is a male, he might prefer a less qualified male candidate over a more qualified female candidate.	Counterfactual checks highlight instances where changing solely the value of race/sex, etc., based variable changes the outcome. In cases where changing the values from out-group to in-group increases positive outcomes, bias is likely and can be explored further.	Training programs utilizing the results of explainable AI checks can be used to make employees more sensitive to in-group biases and shift their judgments from system 1 to system 2 and deliberately correct for the bias.
Anchoring Bias: Here, the user fixates on an initial impression and fails to adequately consider other attributes/factors. The final decision/outcome is skewed due to the initial assessment or anchor. For instance, an individual may form an initial impression on the age of the individual and discount everything else while hiring.	Explainable AI providing relative weights of features considered while making the decision can be used to compare feature weights and identify if one feature dominates the others in an undesirable way. Hence, if age being less than a threshold trumps features like significantly better work experience, and awards and recognition in the field, there could be an anchoring problem. The problem could be present in the opposite direction, with older applicants being removed from consideration, despite other redeeming qualities.	Hindsight exercises, where the human judgment varies from the AI judgment and outlining the differences between relative weights of features considered by humans and AI should make decision makers more aware of decision anchors. Counterfactuals that highlight instances (success stories from a lesser school) should also help.

5.

Mphasis XAI

Industries such as Banking, Finance, Insurance and Healthcare where AI interventions can augment or replace human judgments, are also some of the most heavily regulated ones. Regulatory agencies require greater focus on customer centricity, especially for customers who are vulnerable due to corporate actions. AI interventions in such industries must therefore be designed to ensure fairness of outcome and remove biases that may result in equals being treated as unequal. This assurance, along with its evidence, is critical for AI-based systems to be trusted and adopted by customers.

Considering the need for more trustworthy AI systems, NEXT Labs - the research and innovation wing of Mphasis, has recognized the growing importance of AI model interpretation and identified this as a priority area for the future. We strive to offer explainability features in all critical models through Mphasis XAI - Mphasis' proprietary explainable AI solution that helps remove the black box nature of machine learning model predictions.

Mphasis XAI has been designed to address the mentioned issues around human errors in judgment. It utilizes targeted application of state-of-the-art algorithms to identify and diagnose data and model biases. Mphasis XAI extracts comparative feature importance, displayed as summary plots, which identifies and mitigates anchoring bias by highlighting disproportionately high feature importance. Users also have the option of conducting 'what if' analysis by changing input values.

Mphasis XAI can also, in conjunction with Mphasis' proprietary model drift algorithms, track feature importance over time and identify and mitigate availability bias in response to major news events. The framework's counterfactual analysis can help in identifying representativeness and in-group biases by highlighting how predictions alter with changes in single features like race, sex, etc. Features that should be irrelevant in judgments, but are identified as being significant, would inform organizations about existing biases and help them to take mitigating steps through training and subsequent checks.

Organizations can use Mphasis XAI as a guide to ensure the existing biases are mitigated, and new ones are not introduced when human judgment processes are replaced or augmented by AI solutions. The framework can be utilized for incorporating explainability components from the outset in new AI initiatives, thereby designing trustworthy, interpretable systems that can pass muster on openness and fairness concerns of regulatory bodies and civil society.

6.
Conclusion

Predictable biases and errors in human judgment can have adverse effects on the outcomes for end users. AI systems, due to their reliance on historical data, can perpetuate these errors if they go unnoticed when such systems replace human decision makers. Therefore, identification and mitigation of such judgment biases are vital to ensure appropriate outcomes. Mphasis XAI, which prioritizes human-centric explanations that improve model comprehensibility and interpretability while highlighting areas for improvement in the system, presents a unique opportunity to ensure fairer and unbiased AI systems. It incurs potential resulting benefits such as increased trust among end users, and lower risk of non-compliance with regulatory authorities that are increasingly prioritizing fair and equitable treatment of consumers.

7.
References

[1] Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

[2] https://www.cio.com/article/3274566/the-path-to-explainable-ai.html

[3] European Commission. (2018d): High-Level Expert Group on Artificial Intelligence. [Online] Available from: https://ec.europa.eu/digital-single-market/en/high-level-expert-groupartificial-intelligence

Author

Dr. Ashwani Singh

Manager – Applied AI NEXT Labs

Having experience in analytics and academics, Ashwani is currently a part of the leadership team at Mphasis NEXT Labs. He manages AI/ML to solve business problems.

Ashwani has been part of award-winning projects in the space of emerging technologies. He is also a thought leader in the space of behavioral applications of AI, with a special focus on the emerging area of affective computing.

Ashwani completed his doctoral studies focused on Consumer Cognition and Decision Making, from IIM Bangalore.

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com

For more information, contact: marketinginfo.m@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655

UK
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village
Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

www.mphasis.com

NR 22/09/20 US LETTER BASIL6418