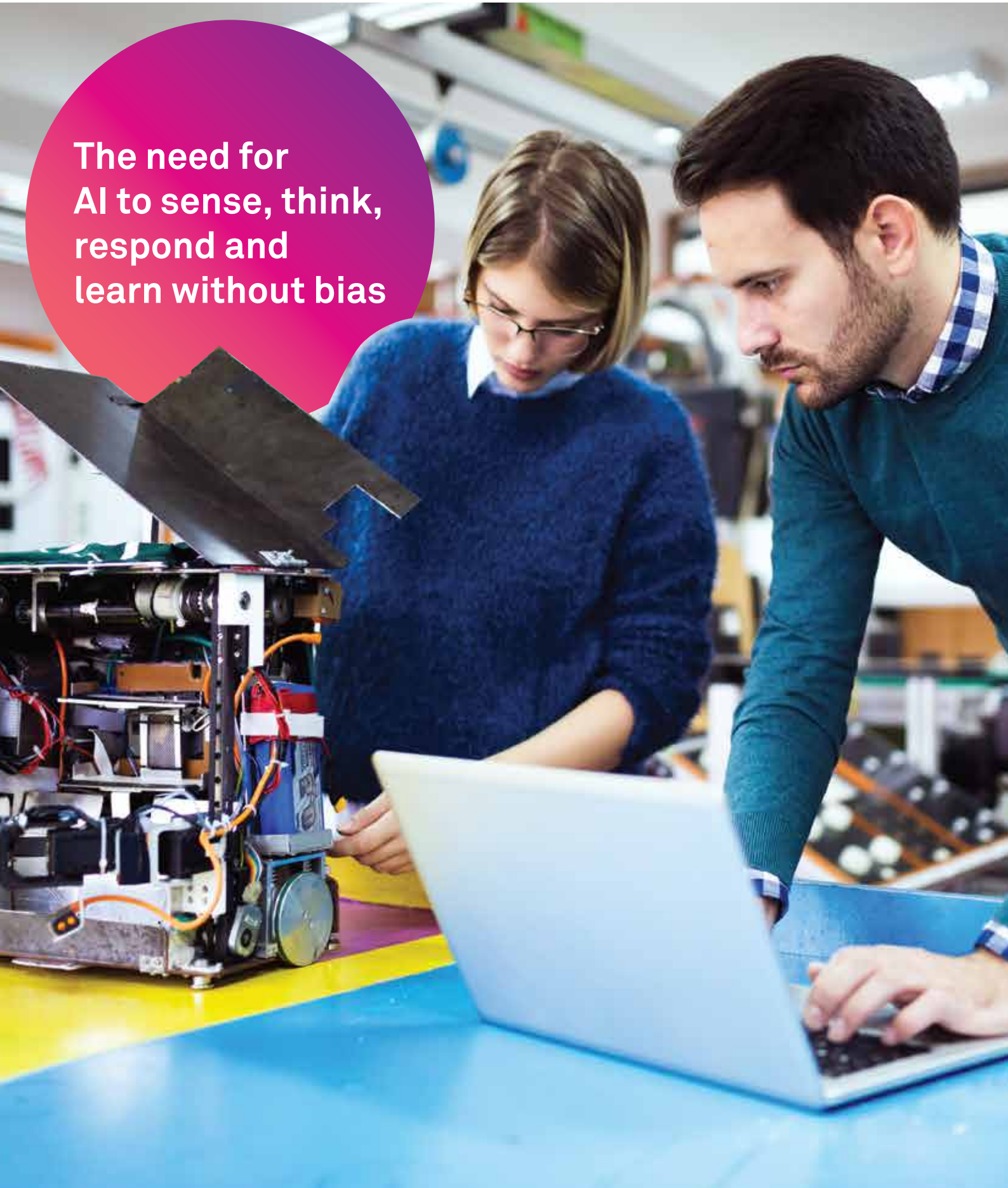
The need for
AI to sense, think,
respond and
learn without bias

Growing adoption of Artificial Intelligence (AI) in systems that assist, rate, and offer advice on how people are treated and what opportunities they are offered is resulting in broad discussions on how to build objective systems. Advanced analytics and AI are being increasingly used to recommend products to consumers, offer pricing tiers to shoppers, manage supply chains and operations, and approve customers to financial products and even automate driving. Therefore, it is essential that the recommendations of machine learning (ML) algorithms are not compromised in any way.

The need for objective AI

Intelligent enterprises leverage AI and analytics to sense, think, respond and learn (Illustrated in Figure 1). It is essential that the ML algorithms that enable this capability are devoid of anomalies.

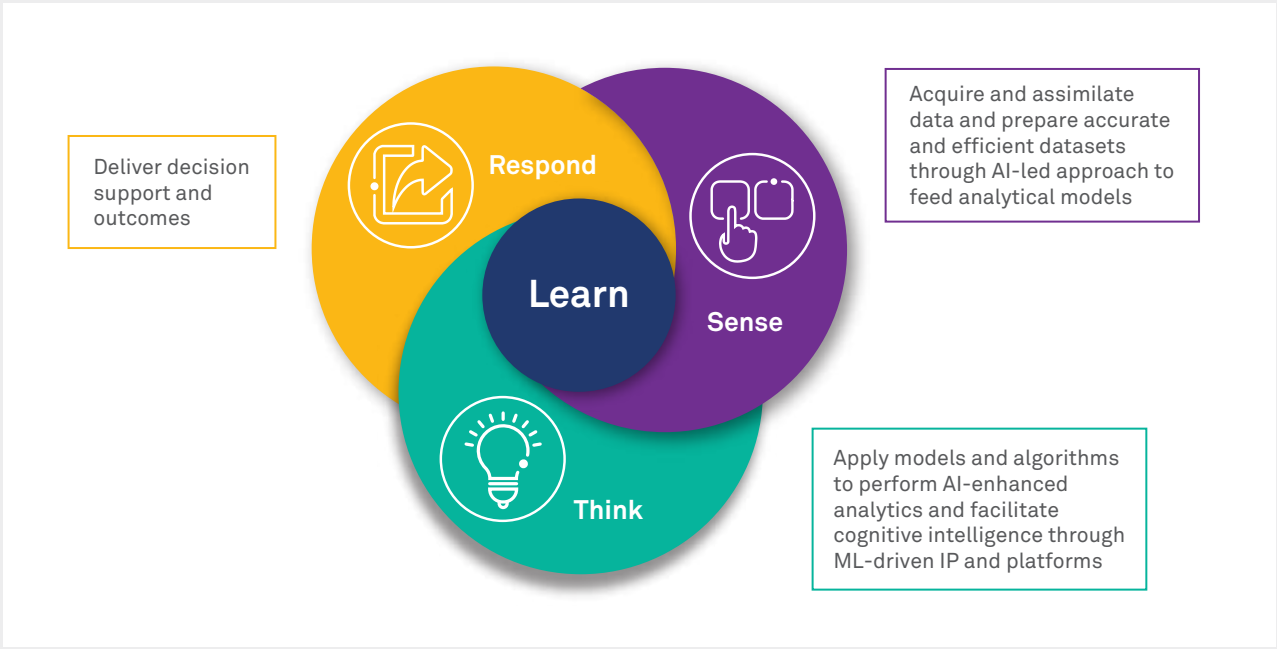

Figure 1: The sense-think-respond-learn paradigm for an intelligent enterprise

Hence, to ensure objectivity and accuracy in data-driven and AI-led insights and decisions, it is essential that ML algorithms are tested multiple times prior to deployment. Here are a few techniques to do this effectively:

- Evaluation techniques using cross-validation and utilizing metrics such as confusion matrices and receiver operating characteristic (ROC) curves (a plot illustrating a binary classifier's performance – a measure of how well it can distinguish true from false) for ascertaining the accuracy of ML algorithms.
- Hardware acceleration of real-time analytics and timely execution to reduce chances of missed opportunity, failure, breakdowns or accidents.
- Structured analytics frameworks and modules that have an in-built mechanism to fault-check and prevent failures and variance in results to make AI systems fail-safe.
- Newer, more accurate algorithms and techniques such as capsule nets that improve upon the many issues with convolutional neural networks (a class of deep, feed-forward artificial neural networks usually applied to analyze visual images).
- Graded validation standards (based on severity of consequences such as DO-178B levels for avionics software, or SAE standards for automotive software) and an ongoing monitoring program for new models prior to deployment.

The uncertainty of how an algorithm would behave under real-world conditions, given its interaction with other predictive algorithms is a critical unknown.

Despite the availability of mechanisms to measure and tune the accuracy based on historic training data, the uncertainty of how an algorithm would behave under real-world conditions is a critical unknown, especially given its interaction with other predictive algorithms that can make it even more complex. Moreover, some of the ML algorithms such as neural networks are akin to black boxes that are impossible to test exhaustively. Ashby's law sounds a clear note of warning here - "If you have complete knowledge of a system, only then it is possible to control it"[1]. How can one control the solution if the system has some hidden properties or your information is incomplete or inaccurate, or if uncertainty abounds about the system's behavior? The business impact, and costs of inaccurate algorithms on human life are too high and hence an ethical framework needs to be setup to remove bias, improve algorithm accuracy, and minimize risks.

Dealing with misuse of AI

The emerging trend of AI and ML algorithms aimed at harming or tarnishing a brand's or a person's image is turning out to be a big social threat. The World Economic Forum considers the viral spread of digital misinformation to be among the biggest threats to human society today[2] - especially with algorithms such as the Generative Adversarial Network (GAN), an image and data generation AI system that performs face and object generation and can even manipulate faces in videos, increasing the scope of misuse.

Widespread adoption of machine learning models requires a clear understanding of the reasons behind predictions. ROC curves, confusion matrices, and error measures are the de-facto industry norms before a ML model is deployed. The Local, Interpretable, Model-Agnostic, Explanations (LIME) is one such framework that explains the predictions of any classifier. LIME works by modifying a single feature and observing its impact on the model output[3]. This helps answer the typical questions a decision-maker might have, such as why was this prediction made or which variables caused the prediction.

For instance, in the case of cognitive software and algorithms that assist doctors in the detection of cancer or other life-threatening illnesses in early stages, the understanding afforded by the LIME framework further provides insights into the model, thereby turning a black box model or prediction into a more traceable and reliable one. This provides the decision maker added confidence when he/she takes a decision based on the algorithm's output.

Tackling machine bias

Bias in a system can lead to a high level of false positives, in turn resulting in low levels of customer satisfaction and acceptance.

Algorithms often exhibit the bias of their creators or the input data fed into them. This machine bias happens when certain hypotheses get eliminated from the hypothesis space or certain hypotheses are preferred over others.

For instance, COMPAS, an ML software used to determine criminal defendants' likelihood to recommit crimes, was biased in how it made predictions. ProPublica found that the algorithm (used by judges extensively in over a dozen US states to make decisions on pre-trial conditions, and sometimes, in actual sentencing) was two times more likely to incorrectly predict that defendants belonging to a particular race were high risk candidates for recommitting a crime[4].

New machine learning techniques called ensemble learning such as bagging, forests etc. make use of the concept of averaging of models and as a result can significantly reduce bias. They are also designed to increase the stability and accuracy of the classification and regression results.

Community groups such as the Algorithmic Justice League, founded by Joy Buolamwini, help promote crowd-sourced reporting and the study of bias in ML and other technologies[5]. Ensuring the involvement of diverse populations in the ethical creation and consumption of ML predictions will lead to further progress in ethics. These developments clearly indicate a distributed, autonomous means of achieving the goals of de-biasing algorithms.

The future of AI

There are ethical and legal consequences of bias in decision making. Machine learning, based on past unfair decisions to a particular race, gender, or sexual orientation, deliver similar biased decisions. Some researchers have gone on to develop a framework for modeling fairness[6].

Then, there are aspects of ethics and even empathy that must be considered by all vendors and users of AI. This could perhaps be addressed by something on the lines of GDPR, perhaps an AIMR - Artificial Intelligence and ML Regulation[7].

AI systems due to their inherent nature may have a trust deficit[8] and this will need to be addressed soon to allay unintended harmful consequences and fears. Beyond specific evaluation techniques for measuring the accuracy of ML algorithms, what we need is hardware acceleration, structured analytics frameworks, and concerted efforts to root out unintended bias and establish objective AI and ML systems.

Reference

[1]https://bit.ly/2FhZBDS

[2]https://bit.ly/2Oh4ot4

[3]https://github.com/marcotcr/lime

[4]https://bit.ly/1XMKh5R

[5]https://www.ajlunited.org/

[6]https://arxiv.org/abs/1703.06856

[7]https://bit.ly/2CvwPyO

[8]https://go.nature.com/2QvBrt6

About the author

Shamit Bagchi

Managing Consultant,
Data, Analytics & AI,
Wipro Limited.

Shamit works as a data scientist specializing in predictive and prescriptive analytics, machine/deep learning and AI consulting. He brings to the table both technical acumen and business consulting expertise based on over 15 years of experience in the big data and software industry, building value propositions for clients in Europe, US and India. He holds a Master of Science degree in Complex Adaptive Systems from Chalmers University, Sweden and an MBA in Marketing & Strategy from the Indian Institute of Management, Bangalore.

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035,
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 175,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**

IND/B&T/MAR-AUG2019