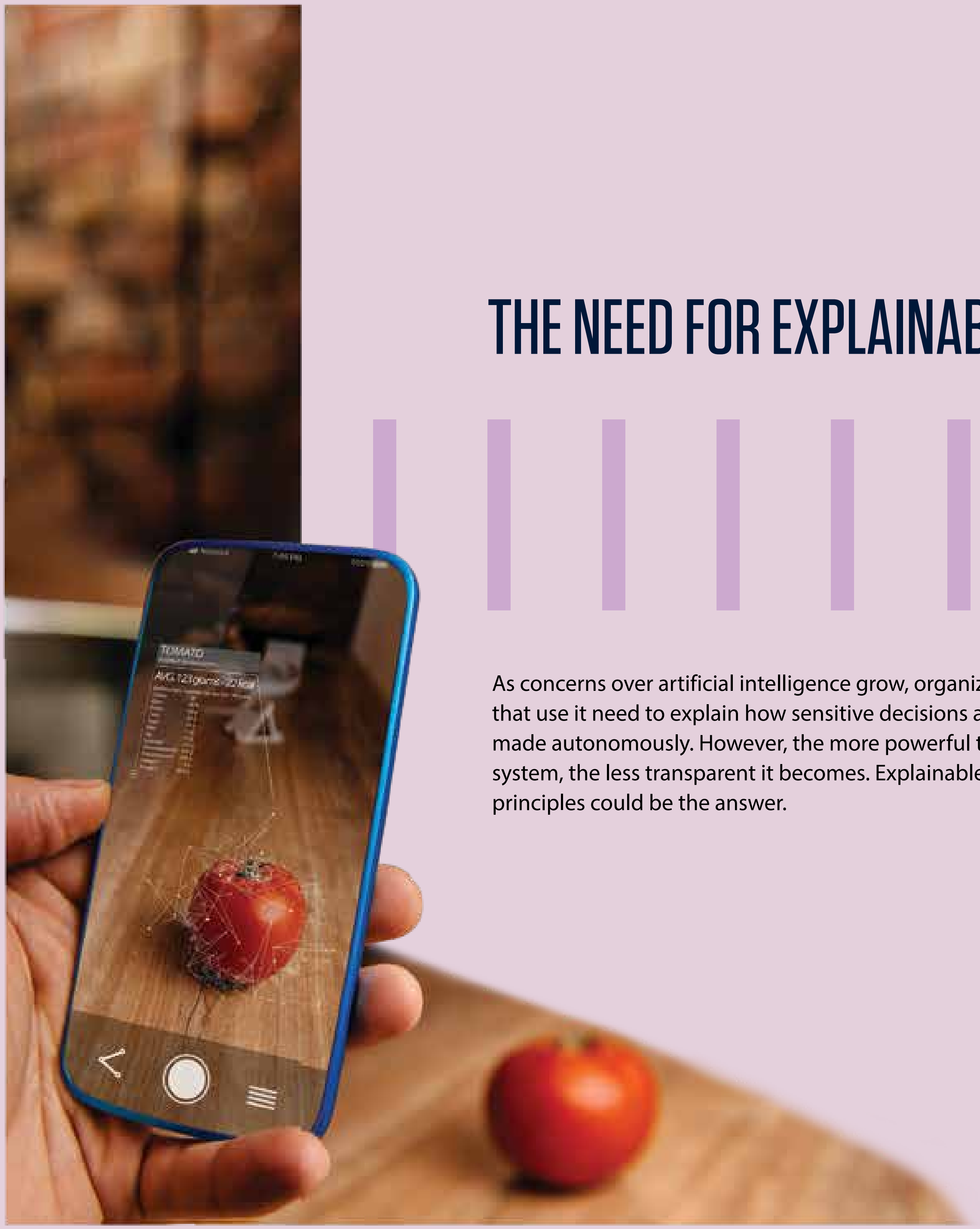
Infosys®
Navigate your next

VIEWPOINT



THE NEED FOR EXPLAINABLE AI

As concerns over artificial intelligence grow, organizations that use it need to explain how sensitive decisions are made autonomously. However, the more powerful the AI system, the less transparent it becomes. Explainable AI principles could be the answer.

Infosys® | Knowledge Institute

Rather than just augmenting human judgment, AI-based systems are now making decisions on their own. Some courts use deep learning to sentence criminals. Banks rely on this technology to grant loans. Transfer learning-based AI can even detect cancer autonomously.

With so much hype, AI is receiving greater scrutiny and criticism about its lack of transparency. How do these systems reach such important decisions? If they can't explain themselves, can we trust them with our lives? These concerns have already led to negative press and litigation. Regulators, official bodies, and users are seeking more transparency in every AI-based decision. In the European Union, the General Data Protection Regulation has enforced the right to AI explanation. Insurance regulators in New York have issued guidance to companies on the use of AI to determine rates and coverage.

Powerful models, opaque decisions

As AI uses ever more powerful algorithms to reach its decisions, the ability to understand the logic becomes increasingly difficult. Geoffrey Hinton, a University of Toronto computer scientist known as the "godfather of deep learning," sums up the problem nicely: "A deep learning system doesn't have any explanatory power. The more powerful the deep learning system becomes, the more opaque it is."[1]

When algorithms make incorrect assumptions, the opaqueness makes it more difficult to determine what went wrong. One example is an AI classifier used to distinguish wolves from huskies. If the training data showed pictures of wolves in winter settings, it might be biased toward snowy conditions. A misguided algorithm would learn to use snow as a feature for predicting wolves. Given new data without snow, the classifier might determine that the animal was in fact a husky even when wolf was the correct answer.

> The more powerful the deep learning system, the more opaque it becomes

Misclassifying animal species creates one set of problems. However, categorizing people can lead to more severe reputational consequences. AI bias has already found its way into issues of race and gender. ImageNet, a large visual database, announced it would remove more than half a million images from its records after finding racial bias. AI classifiers often decided that white women were "stunners," whereas other races were given far less flattering monikers.[2]

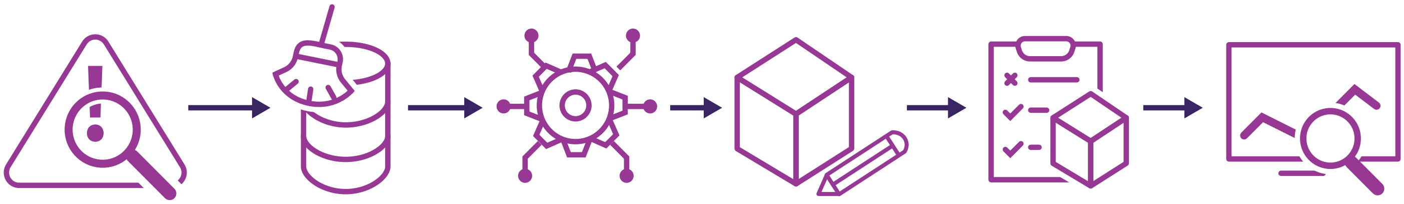
In health care, AI needs to be even more rigorous. Three U.S. universities collected and trained chest X-ray data on more than 150,000 patients to detect signs of pneumonia. The convolutional neural network performing the analysis was far less accurate in real-world usage because the network cheated on the predictive task by first rating the quality of the X-ray machine providing the test data.[3]

To be trusted, AI must not only classify objects correctly but also explain its logic. For the wolf-husky classifier, the model should be able to highlight the specific parts of the animal that led to its decision. This is known in industry parlance as XAI.

XAI workflow

To ensure that XAI is effective, the technology must be used throughout the AI life cycle, from data cleansing to model creation to monitoring (Figure 1).

Figure 1. XAI must be used throughout the AI life cycle



Problem selection-definition
- Selection of an unbiased problem

Data cleansing, data selection
- Data cleansing
- Check for bias class representation
- Oversampling, undersampling

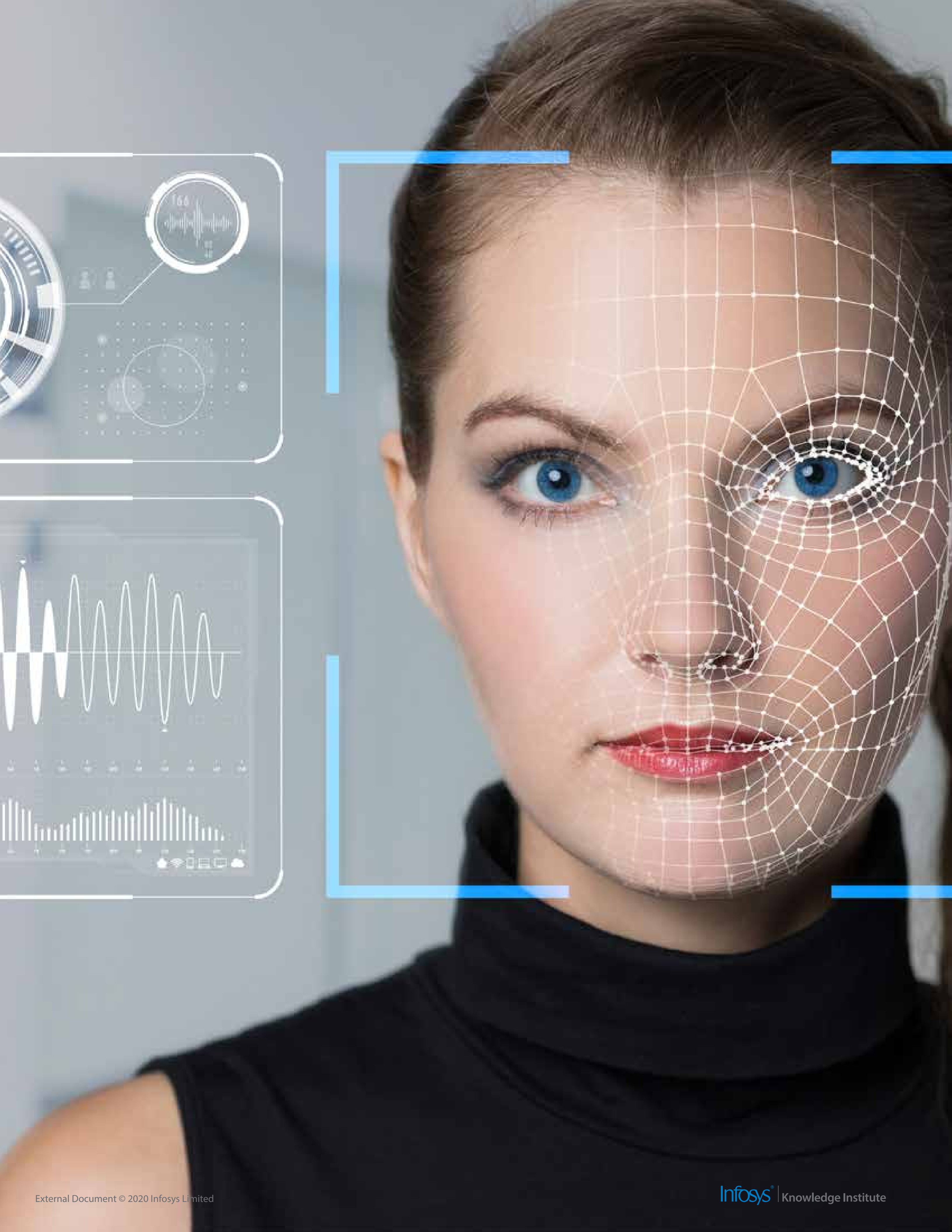
Algorithm selection
- Classical vs. deep learning

Model creation
- Feature selection
- Weights
- Hyperparameters

Model training and testing
- Data distribution in training and testing
- Accuracy metrics: F-score, precision, recall.

Monitoring
- Reproducible results
- Verifiable
- Identify new data elements
- Manual sample reviews

Source: iCETS



In the AI workflow, XAI requires the following:

- Data must have equal representation for all classes and is checked for bias.
- The correct algorithm is used for training and testing. This could mean choosing a classical algorithm, such as logistic regression, over a convolutional neural network or other fancier networks.
- The model must have the right features and give appropriate weights to each one.
- Training and testing require agnostic data verification using frameworks such as local interpretable model (LIME) and SHapley Additive exPlanations (SHAP). Training and testing should go beyond mere accuracy, using metrics such as F-score, precision, and recall.
- The AI model should be monitored and results verified by keeping tabs on incoming data varieties.

Principles for good XAI

Process is important, but companies also need a strong set of principles to guide AI implementation. These include:

- Bias detection — Firms should make sure datasets are unbiased and nondiscriminatory, with attributes weighted correctly and used with discretion. In a dataset used to generate credit scores, age can be tested for bias by creating binary variables. For example, ages greater than 25 are set to 1 (privileged group), while an age of less than 25 is set to 0 (unprivileged group). A mean difference between favorable results for each group can then be calculated. A score of .16 indicates that the privileged group has 16% more positive outcomes. This means the data must be reweighed, where weights of individual samples are changed to balance the dataset before it is fed into the machine learning model.
- Human involvement — The output must be easily understood by humans, and people must be kept in the loop at all times. This is particularly important in fraud detection and law enforcement uses.
- Justification — To be true XAI, predictions made by the system must be justifiable. This means transparency in the use of feature data. This helps understand how the model is thinking or which features of a given output were emphasized by the model.
- Reproducibility — The model should be consistent when making predictions and remain stable when testing the system with new datasets.

Explainability by justification

Explainability by justification is among the most important elements AI will need going forward. This requires highlighting features that contribute to the accurate prediction of a class rather than a random feature such as snow, as demonstrated in our wolf-husky classifier.

> AI will need to use 'explainability by justification' going forward

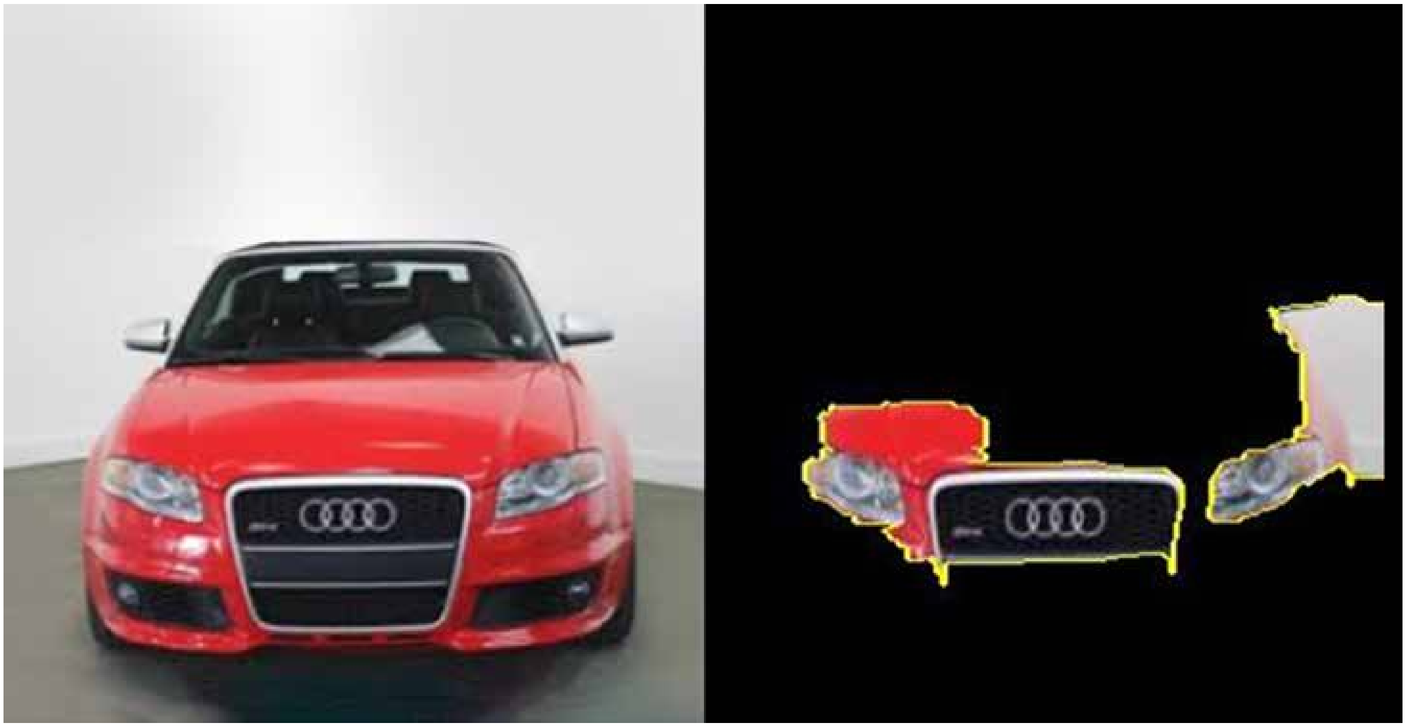
When implementing AI, there are explainability by justification models. The LIME model highlights the parts of an image that are dominant in the prediction of a class. For both image and text classification scenarios, the SHAP model is available. This gives insights into how a certain layer in the deep learning model is affecting output probabilities.

LIME

In LIME, a temporary model is trained to mimic deep learning black-box predictions. Given a sample input, the temporary model generates an interpretable output dataset. It does this by creating various permutations of the given sample (and their corresponding output) and trains a simple and more interpretable local model on this dataset. The output of LIME is a list of explanations for the model arriving at a certain decision, showing the contribution of each feature to the prediction of a data sample.

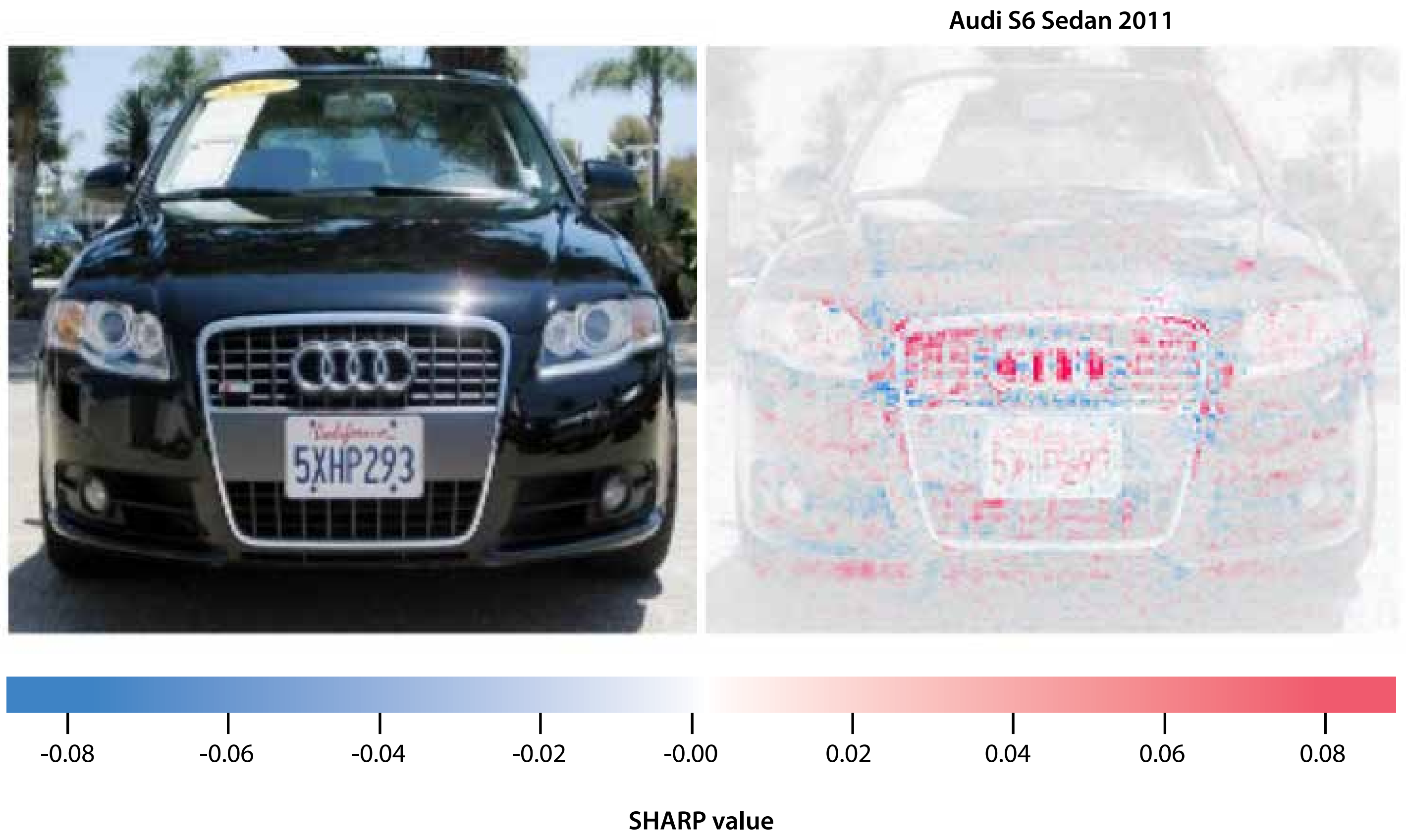
At the Infosys Center for Emerging Technology Solutions (iCETS), a model was trained to classify cars based on their visual features. A transfer learning paradigm was used for model training (ResNet architecture with pretrained weights). The last layer in the deep learning algorithm was trained on nearly 200 car categories — each containing 50 images — with an accuracy of 90% on test data. To bring LIME into the equation, the car classifier passed its images to a LIME library to verify what regions the model was using for classification. LIME created 1,000 samples out of the image by trying various permutations of superpixels (a collection of similar pixels) based on segmentations. In the case of an Audi (Figure 2), the output explainer showed that the model focused on the Audi logo to achieve its 90% success rate.

Figure 2. The LIME model showed that the logo was the most important feature



Source: iCETS

Figure 3. With SHAP, the red pixels contribute positively, whereas the blue pixels are negative



SHAP

The SHAP library is primarily based on game theory. The contribution of each feature to the prediction is calculated. With images, the features can be pixels or superpixels, and their contributions can be either positive or negative. First, an average prediction capability is found using a sample dataset. Then, the contributions of individual features are calculated by giving different permutations to the model and calculating whether the feature increases the prediction capability or reduces it. SHAP is powerful in that it gives humans a layerwise explanation of deep learning models.

iCETS used the same car classifier model for explanations using SHAP (Figure 3). A layerwise explanation was found using a pretrained model, an image, and a background dataset.

The power of both LIME and SHAP is that we know not only what is being predicted but also why. If the model is not picking up the correct features for prediction, the model can be fine-tuned while also making complex machine learning algorithms and models more transparent and trustworthy.

> With LIME and SHAP, not only do we get good predictions, but we find out why they were made

The future of AI

As XAI generates greater interest, domain experts are coming together to lay foundational principles that machine learning and AI models should follow. More complex machine learning models are under the microscope, given their lower levels of transparency.

Google, after setting a goal to be AI-first in 2017, is pioneering XAI by integrating a What-If Tool in its proprietary TensorFlow framework for deep learning. In this way, Google is looking to be the torchbearer for making AI less mysterious by offering XAI-as-a-service. "Explainable AI at Google is a set of tools and frameworks to help you develop interpretable and inclusive machine learning models and deploy them with confidence," said AI researcher Andrew Moore at a recent Google conference. "With it, you can understand feature attributions in AutoML Tables and AI Platform and visually investigate model behavior using the What-If Tool."[4]

Further afield, The Institute for Ethical AI & Machine Learning is currently creating a framework that ensures ethical and conscientious development of AI projects across all industries. In its work toward this, the institute has published ethical AI principles and also developed an open source GitHub toolbox for explainability.[5]

AI could go in one of two ways. The first future is one where businesses could implement XAI based on principles laid out in this paper and garner greater trust from the public and government. The other future is one where regulatory agencies comb through training data for stereotypes to ensure that AI decision-making is fair and justified — a worrisome scenario for corporations. This accelerated adoption of XAI in business is likely to ensure that Gartner's prediction of a $4 trillion global AI economy by 2022 will come to pass.

References

[1] Explainable Artificial Intelligence (AI), Mike Ridley, Feb. 5, 2018, Open Shelf.

[2] 600,000 Images Removed from AI Database After Art Project Exposes Racist Bias, Zachary Small, Sept. 23, 2019, Hyperallergic.com.

[3] AI tools may fail during key medical diagnosis: Researchers, Nov. 10, 2018, Hindustan Times.

[4] Google's new 'Explainable AI' (xAI) service, Tirthajyoti Sarkar, Nov. 25, 2019, Towards Data Science.

[5] EthicalML/xai, GitHub.

Authors

Sudhanshu Hate

Senior Principal – Infosys
sudhanshu_hate@infosys.com

Ram Swaroop Mishra

Data Science Engineer – Infosys iCETS
ram.mishra@infosys.com

Producers

Jeff Mosier

Senior Consultant – Infosys Knowledge Institute
jeff.mosier@infosys.com

Harry Keir Hughes

Senior Consultant – Infosys Knowledge Institute
harrykeir.hughes@infosys.com

Anu Mary Tom

Senior Consultant – Infosys
anumary.tom@infosys.com

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision making on critical business and technology issues.
To view our research, visit Infosys Knowledge Institute at infosys.com/IKI

For more information, contact askus@infosys.com

Infosys®
Navigate your next