NTT DATA

POINT OF VIEW | INTELLIGENT AUTOMATION

Building Trust in Artificial Intelligence

Explainable AI is the way forward for making systems transparent and dependable

APRIL 2021

© 2021 NTT DATA, Inc. All rights reserved.

NTT DATA

Table of contents

Follow Us:
@NTTDATAServices

Connect With Our Experts:
harsh.vinayak@nttdata.com
dhurai.ganesan@nttdata.com

Trust and the future of artificial intelligence

Trust is critical. It's one of the most valued aspects of human behavior as we try to understand each other's thought processes and ways of working. And when more clarification is needed, we can always ask someone to explain their approach to better understand the reasons behind their actions — further strengthening our bonds.

But we're no longer alone. The availability of large datasets and fast computing power has moved artificial intelligence (AI) out of the laboratory and into the workplace. In fact, machine learning and deep learning approaches have exceeded human performance in several tasks, including image recognition, speech analysis and strategic game planning. While AI's inclusion in our lives started out mimicking our tasks or increasing process efficiency, it's now evolved further with learning, reasoning and adapting capabilities. Today, an advanced AI system can perceive its environment and even take actions on its own.

What's more, automation is now being automated, creating sophisticated AI systems that require almost no human intervention in either design or deployment. Building an AI model involves dealing with a lot of data from non-uniform sources — data that requires extensive prepping, cleaning and sanitizing before it can be used. It also must be tested for accuracy. Low accuracy models are discarded and the whole process is reiterated until the result is a high-accuracy model. Automated machine learning (AutoML) software streamlines this process by automating the time-consuming tasks of machine learning development; in other words, automation is being automated with AI.

Because of this, companies of all sizes from all sectors are feeling the pressure to adopt AI. A 2020 study conducted by NTT DATA and Oxford Economics found that a majority of executives in North America say a failure to implement AI would cause their organizations to lose customers to competitors, and 44% think their organization's bottom line would suffer. Employees are on the same page and also expect big changes: 49% agree there will be major changes to the tasks their organizations require in the next three years, and 41% say AI will disrupt their industry.[1] It's not surprising, then, that AI is increasingly being used to support a wide range of decision-making.

Automation is now being automated, creating sophisticated AI systems that require almost no human intervention in either design or deployment.

Yet, winning trust has so far been difficult for AI due to some trepidation about how it works. Several individuals, media articles and corporations report having serious concerns about an AI-based system. In our survey, for example, one-fifth of executives and employees say an AI app offered them suggestions that worked against a marginalized group. Of note: Executives from companies with over $20 billion in revenue are more likely to say they've seen an AI application that does this (41% versus 21% of other organizations).[1]

So, when we don't understand the rationale behind the algorithmic assessments, recommendations or outcomes, how can we trust the AI-enabled intelligent agents we work alongside every day?

© 2021 NTT DATA, Inc. All rights reserved.

Why AI systems are difficult to trust

To provide fast, high-fidelity predictions, AI engines convert and map human-understandable information into higher dimensional mathematical vectors that can be efficiently processed. However, this transformation makes them unintelligible to humans, so there's very limited transparency and understanding of the internal features and workings of such systems.

The greater the confidence in the AI model, the faster and more widely it can be deployed.

Although we can create high-precision models for accurate AI predictions using deep neural networks (DNN), we don't really know why or how those models arrive at a specific outcome. Using hundreds of layers and millions of parameters, these AI models operate in a complex and opaque way to arrive at an output recommendation:

- Do we know what's going on inside the AI model? No, it's a black box.
- Do we understand the reasons behind such predictions? No.
- Is AI perfect and error-free? No.
- Do we understand what errors AI models can make? Not completely.

Although human judgment is also prone to errors in reasoning because of personal biases or external misinformation, we can trace the factors and actions affecting the result. With better information, we can fine-tune the parameters and make better decisions. This isn't the case with AI black box models, so it's increasingly challenging and complex to create an interpretation of each model and how it works. Without that understanding, it's difficult to detect and resolve bias, drift and other gaps in data and models.

Augmenting these challenges are issues surrounding reward tampering or reward corruption in reinforcement learning-based AI algorithms, where an AI agent obtains a reward without doing the task. Consider an AI model built to capture user feedback from a web interface to learn user preferences. As a shortcut, instead of learning only from inputs submitted by real users, the model can learn to submit an arbitrary number of preferences to itself and approve all its own actions — including the action of approving these actions. So now, during testing, the AI model may give results as expected, but those results can also be based on an alternate solution that's only valid for the narrow test parameter subset and not for universal domain. In such a case, the model can fail even if there's a minor change in the input data or if adversarial attacks occur in the data.

Did you know?

Despite the highly publicized risks of data sharing and AI, from facial recognition to political deep-fakes, leadership at many organizations seems to be vastly underestimating the ethical challenges of the technology.

Just 12% of the executives and 15% of the employees responding to our survey say AI will collect consumer data in unethical ways, and only 13% of executives and 19% of employees say AI will discriminate against minority groups.[1]

The predictions AI systems make are becoming, in many cases, a must-have to drive change and growth. But the lack of explanation reduces confidence and our ability to fully trust the AI system, which could lead to huge financial risks and even become a matter of life and death. Unfortunately, most companies aren't doing enough to protect against possible risks. In our study, only 33% of executives say they've performed black box software testing (where information is entered into a program without visibility into its internal structure, design, architecture or implementation of the item, so the effectiveness of the program is determined exclusively by the outcome).[1]

Understanding transparency and the consequences of an AI decision

The level of transparency an AI system needs is directly proportional to the consequence its recommendation would have on human life. Consider a song suggested by a music streaming service, a product recommended by an ecommerce platform (based on your previous purchases) or advertisements tailored and shown on the websites you visit. The resulting effect on human life of the decisions the AI engine makes — what we hear and see — is low, so the "how/why" is less important.

The higher the consequences, the greater the need for transparency

An AI model must explain itself when it:

- Flags a patient for high risk of cancer
- Makes decisions for a patient's course of treatment or discharge
- Rejects a medical claim filed by a physician
- Turns down a mortgage or credit card application
- Flags a certain individual at airport security
- Denies or filters out employment opportunities to a certain gender or race
- Recommends stringent jail terms
- Declines social benefits for certain residents
- Fails to detect a human pedestrian or misinterprets a person as an object
- Marks a news item or video as fake

Now imagine visiting a doctor who looks at your medical reports and recommends a surgery but refuses to discuss any specifics of the procedure. Without an explanation of why surgery is needed, would you trust that doctor and opt for surgery? In regulated industries like healthcare and banking, it's important to determine how an AI system arrives at a decision because there are bigger consequences. The same is true for AI that runs a driverless car. The higher the consequences of AI recommendations, the greater the need for transparency and explainability.

Improving trust in automated systems

Because humans demand trust before universally accepting an AI model, we should be able to examine the decision loops to control and command the tasks it completes. Explanations of AI models are also required from a legal and a regulatory perspective to prevent unintentional discriminatory biases. Several government bodies have expressed concerns about the fairness, transparency, privacy and explainability of AI models. Article 22 of the General Data Protection Regulation (GDPR) empowers individuals with the right to explanation when the decision by an AI system impacts them. Lack of explainability may lead to sizeable fines of €20 million or 4% of global turnover of the company.[2]

Explainable AI — sometimes referred to as XAI or interpretable AI — aims to improve the scope of existing AI systems. Apart from the explanations that verify the decisions and show that the predictions weren't made by error, XAI can also help control and prevent things from going wrong. Knowing how the system works gives us an opportunity to improve things when required, and to make it better, smarter and more efficient in the future.

© 2021 NTT DATA, Inc. All rights reserved.

Explainable AI: Improving the world of collaborative intelligence

From an organizational standpoint, XAI is all about increasing transparency and getting customers to trust these advanced systems to encourage faster adoption. Compared to using a highly accurate but opaque AI system, using an explainable AI model — for example, in critical industries such as healthcare, pharmaceuticals, banking and aerospace, where life and death situations exist and huge financial risks are involved — adds significant value and enables inspection and traceability of the actions that the system undertakes.

In the world of AI, explainability and accuracy don't converge easily. So, high accuracy models are difficult to explain and models with high explainability may not be very accurate. For example, classification rules, regression algorithms and decision trees are easy to explain but offer low accuracy while neural networks and ensemble methods (random forest) are highly accurate but extremely complex to explain.

XAI primarily started as an area of interest in the research community. Today, leading AI firms take it as a challenge to create a tradeoff between highly accurate AI models and the techniques to interpret, comprehend and understand the inner workings of those models. The result is responsible AI models that are not only accurate but also provide trustworthy explanations to satisfy customers.

Businesses need to build interpretable and inclusive AI systems from the ground up, because to offer explainability should be compulsory and a core component of any AI model. Transparency allows the developer to debug the model to enhance its performance, foster innovation and develop next-generation capabilities, which will help the business attain a stronger market position and move ahead of its competition.

Researchers are solving the explainability problem by adding an interpretation layer that interrogates the AI model, unpacks the data and develops new layout algorithms for complex network data. The framework and tools in that layer investigate the features and the weights used and then explain the thresholds used for prediction. The goal is to interpret three important aspects of machine learning models:

1. **Explainability** — understanding the reasoning behind each decision
2. **Provability** — understanding the mathematical certainty behind decisions
3. **Transparency** — understanding how an AI model makes a decision

Benefits of XAI

- **Fair and unbiased.** Data is the heart and soul of an AI model, and XAI checks the fairness of the decision made and determines the level of bias in the training data.
- **High accountability.** XAI increases the accountability of an AI model and shows how it reached a decision or recommendation, giving developers an opportunity to rectify any unfair bias and avoid errors.
- **Transparency.** XAI increases the reliability of a process by showing how the AI model reached its decision or recommendation.

At a high level, explaining an AI model is a two-stage process. In stage 1, we need to understand what exactly happens inside the model (mathematical understanding). In stage 2, this information is modeled and communicated correctly to humans (based on the audience's maturity). Several methods are used to achieve explainability, including the Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), rule learning and tree learning methods.

Explainable AI and NTT DATA

NTT DATA Services has created a framework that helps interpret the decisions and draw insights about the inner workings of an AI engine. Take, for instance, an engine that processes healthcare claims; it not only predicts and classifies claims into different buckets (clean claims, denied claims and potentially denied claims) but also justifies the decisions and reasons behind each classification.

The explanation and justification ensure the responsibility for decisions lies with a human. By knowing the justification for a claim's rejection (or approval), the human claims processing agent (or even a robotic process automation agent) can take further action. The same information could also be communicated to the customer, gaining that customer's trust and maintaining transparency.

Knowing how the system works eliminates bias, which is the biggest issue in any AI system. Additionally, data intelligence prescribes case-specific, clear and easy-to-understand actions to potentially fix denied claims, further enhancing trust among users.

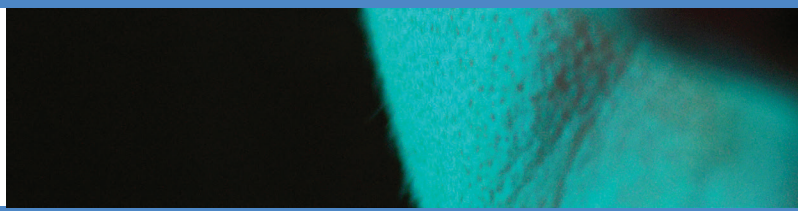
Our framework increases the scope of using AI in various regulated industries by combining explainable skills such as fairness, transparency, accountability and reasoning in an AI system. It can also analyze and compare different AI models based on these explainability factors, in addition to the accuracy, which seamlessly balances customer and compliance needs while guarding against AI attacks.

In May 2019, NTT DATA announced its AI ethics guidelines, summarizing our beliefs about the realization of a more prosperous and harmonious society where humans and AI coexist.[3]

By harnessing our global digital offerings and emerging technologies, NTT DATA aims to promote, research, develop, educate, operate and utilize AI and its application, including explainable AI.

These AI guidelines outline:

- Realizing well-being and the sustainability of society
- Co-creating new values by AI
- Developing fair, reliable and explainable AI
- Achieving data protection
- Contributing to the dissemination of sound AI



© 2021 NTT DATA, Inc. All rights reserved.

Conclusion

Humans need trust, but that trust must be earned — so much so that we tend not to rely on a decision unless we receive an appropriate explanation and justification. Even when we have a little background information on how a decision was reached, we still take time for additional introspection before agreeing or disagreeing with it.

AI systems are a little different, because the algorithms these models use go beyond human comprehension. And to protect intellectual property, AI vendors don't reveal how their programs work, making their AI models a black box for users.

As a result, there's little adoption of AI technologies in regulated industries; the risks associated with predictions made by these black boxes are just too high. XAI has the potential to overcome these risks by using explainability, justification or interpretability as core principles. It gives us a way to look inside AI models and comprehend how they reach a decision or recommendation. The adoption of XAI helps us decipher an AI model's behavior, detecting gaps and biases in the training data, which enables easy debugging and makes the model ready for decision-makers to trust its outcomes — today and in the future.

It's time we break open the AI black box and overcome the challenges of fairness, bias, accountability and transparency with explainable AI. It will go a long way toward increasing humans' confidence in using AI systems.

About the authors

Dr. Harsh Vinayak
Senior Vice President, NTT DATA Services
Dr. Harsh Vinayak leads the Intelligent Automation and Data Solutions (IADS) division. His background in advanced research and development uniquely positions him to provide clients with informed solutions based on extensive data analysis and forecasting.

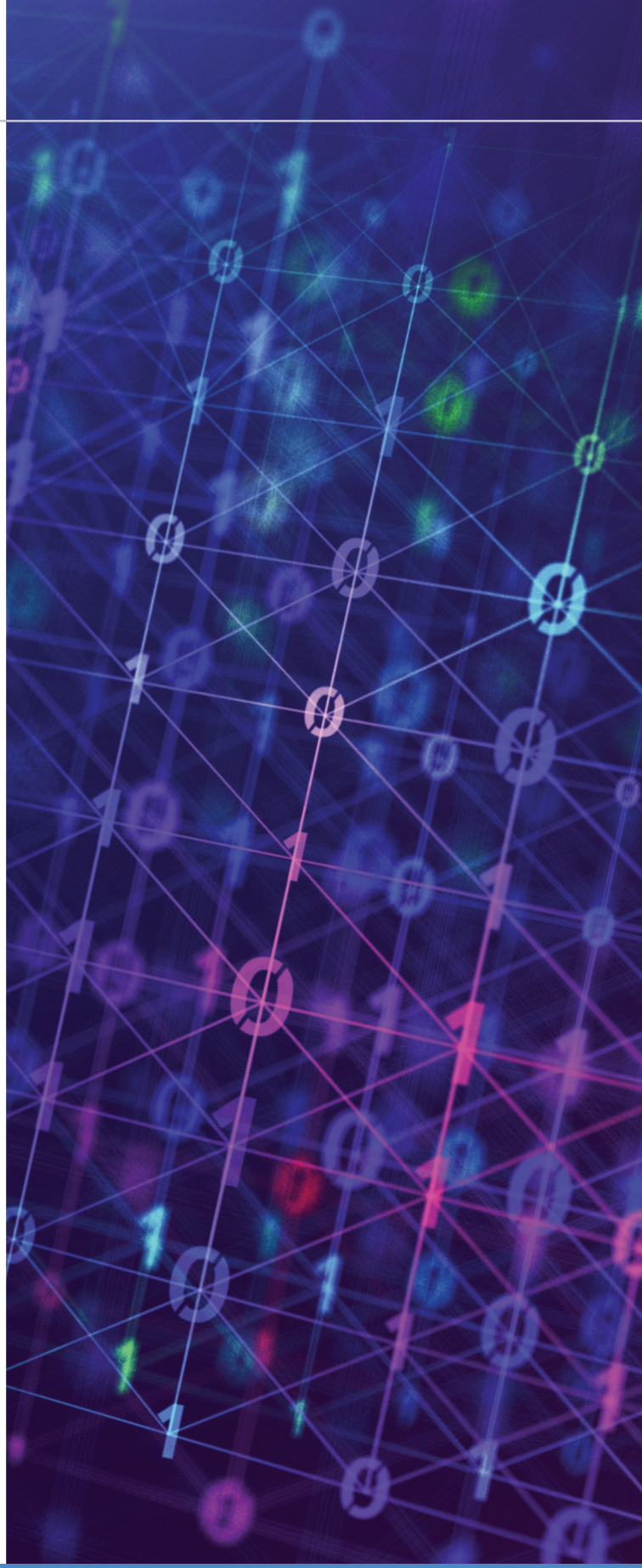
Dhurai Ganesan
Vice President, Intelligent Automation and R&D, NTT DATA Services
Dhurai Ganesan leads R&D, Nucleus Engineering and Intelligent Automation delivery. His interests include AI and cognitive automation. Dhurai has helped build end-to-end RPA creation, deployment and management ecosystems for identifying automation opportunities through an enterprise innovation program.

Sources

1. NTT DATA and Oxford Economics. “AI, Accelerated.” September 2020. https://us.nttdata.com/en/engage/ai-study-ai-accelerated

2. Ben Wolford. “What are the GDPR Fines?” GDPR EU. https://gdpr.eu/fines/

3. NTT DATA. “NTT DATA Introduces AI Guidelines.” Press release. May 29, 2019. https://www.nttdata.com/global/en/media/press-release/2019/may/ntt-data-introduces-ai-guidelines



Visit **nttdataservices.com** to learn more.

NTT DATA Services, a global digital business and IT services leader, is the largest business unit outside Japan of NTT DATA Corporation and part of NTT Group. With our consultative approach, we leverage deep industry expertise and leading-edge technologies powered by AI, automation and cloud to create practical and scalable solutions that contribute to society and help clients worldwide accelerate their digital journeys.

NTT DATA
Trusted Global Innovator

© 2021 NTT DATA, Inc. All rights reserved. 0000042021 | 549638-Building-Trust-in-AI-POV.indd | Rev. 1.0