An Introduction to XAI

Explainable Artificial Intelligence

Excella

AGILE | DATA | DIGITAL | MODERNIZATION

Introduction

In a world fueled by digital data, the use of artificial intelligence is prolific – from the automation of human processes to discovering hidden insights at scale and speed. Machines can do many tasks far more efficiently and reliably than humans, resulting in everyday life increasingly resembling science fiction. This inevitably sparks concern about controls – or lack thereof – to inspect and ensure these advanced technologies are used responsibly. Consumers want reassurance about ethical use and fairness related to AI. Businesses need to mitigate risk of unintended consequences when employing these advanced, complex solutions. Enter: Explainable AI, an attempt to create transparency in the "black box" of artificial intelligence.

Continue reading to learn more about:

What is XAI?

Explainable AI (or XAI) is an emerging area attempting to focus on increasing the transparency of AI processes. Forbes recently wrote about the movement, observing:

> As humans, we must be able to fully understand how decisions are being made so that we can trust the decisions of AI systems.

XAI aims to inspect and reflect the steps a machine took when making a decision or taking an action.

Artificial Intelligence is essentially incredibly complex math.

With the advances in computing power and the scalable compute cloud vendors offer, the ability to run highly advanced mathematical models in a cost-efficient manner is now mainstream. This has powered the rise of machine learning models – algorithms that can learn and adapt without human intervention. Previously, this type of compute intensive calculation required a supercomputer (think the original IBM Watson). Now businesses large or small can employ machine learning models to advance their missions. New business models have even evolved due to machine learning advancements that would not have been previously viable (e.g. Uber, Waze).

While there are many benefits to consumers and businesses alike when leveraging machine learning, the complexity of these AI models can make it difficult to answer common questions, like:

- Why did the model make a decision or prediction?
- When the result is unexpected – why did the model pick an alternate choice?
- How much confidence should be placed in the model results?

To be able to regulate (informally or formally) AI, we must be able to understand and explain it.

The need for transparency

The expanding use of even more complex models using deep learning methods for use cases such as facial recognition, voice to text, or TV show recommendations adds urgency to open the AI ‘black box’ and provide transparency.

In addition, efforts by the European Union (EU) are formally defining what “trustworthy AI” is with Transparency defined as one of seven key requirements. The EU led the way, introducing sweeping data protection regulations with the introduction of GDPR and defining the same level of protections for its citizens with AI.

Making the Case for XAI

What exactly is explainability in the world of AI? Think of it as a two-step process – first, interpretability, the ability to interpret an AI model, second, explainability, to be able to explain it in a way humans can comprehend.

Interpretability

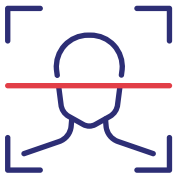
The extent to which an AI model's decisions can be comprehended (the raw mechanics of the model workings) and anticipated (we know the likely model output).

Explainability

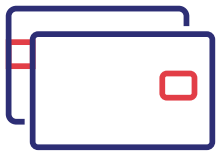
Going several steps further, where a model's decision or prediction can be made transparent and communicated to humans. In other words, a human would be able to understand why a decision was made without being an expert in advanced math.

Understanding the model mechanics **(interpretability)** is necessary first to support the second step, translating this for human comprehension **(explainability)**.

AI transparency is important to identify where models are operating with negative consequences. For example:



Facial recognition that discriminates - COMPAS software is likely racially profiling when forecasting likelihood of reoffending

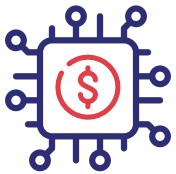


Credit cards with lower limits because of gender - Apple's credit card was approving lower credit limits for women

These problems can be addressed, or their impact mitigated if XAI was part of the requirements as an AI model was developed. By understanding how a model reached its decisions, we can audit its fairness across different demographic groups and ensure protected groups are not discriminated against or ignored.

XAI Considerations

Introducing XAI requires review of the possible downstream impacts.



Initial Investment

Using an XAI approach to AI model development may increase the amount of initial investment to support model transparency requirements. It could also deter the selection of an advanced technique that provides superior results, but cannot be easily explained. Alternatively, the potential risk of unintended, negative outcomes using a 'black box' AI approach could result in much larger costs in the long term to remediate.

Intellectual Property

Some of the latest research has shown that as the explainability of a machine learning model increases, the security of that model can decrease. Given full explainability, a model can be reverse-engineered or recreated. The recreation of the model can pose a threat to the intellectual property of the technologies involved in its original creation. Additionally, such information can expose the model to hacking or hijacking. Mitigating this risk could be the ability to explain the model and make it transparent is possible, but this level of information is only shared with a limited audience.

Unintended Data Consequences

The ability to explain and understand how complex AI models operate is one component of addressing AI ethics. Alone, XAI is not enough. The data used to fuel the model is another key factor.

Unintended bias in AI frequently occurs because of the data sets used to develop and train an AI model. Data "teaches" or trains the model to respond based on data patterns. Using unreliable, inconsistent data means the AI model will likely have less predictable or dependable results. Likewise, limiting the data population employed during model development and training can lead to a model with a confined view of possible scenarios and outcomes, increasing the risk of unexpected results once live. To be effective, a model needs to have enough examples to train and learn. The data also needs to be representative – having a sufficient volume of examples to include all possible user populations. Time and care when selecting training data sets, with bias prevention in mind, will help in the long run.

A Use Case for Understanding AI Models and the Importance of XAI

Grants Analytics Portal

Health and Human Services, Office of Inspector General (HHS OIG) is the largest grant-making organization in the federal government. It's important for HHS OIG to use analytics to better understand our grant profile and to target our resources related to grant fraud, waste, and abuse. The HHS OIG grants analytics portal (GAP) is a data fusion project that brings several grants data sets together and uses analytics to bring the most interesting data forward for OIG agents, evaluators, and auditors.

Text Analytics in the GAP

The GAP uses text analytics to make millions of pages of documents accessible to OIG staff. For example, the team uses text analytics to turn the A-133 single audit reports into machine readable text. Rather than reading through more than 30,000 documents a year, OIG staff has a database they can query. OIG has also used text data to mine key grant documents for identifiers that can be used to link several different data sets together.

Deployed Neural Network Models in the GAP

The A-133 single audits contain key findings related to the internal controls of certain grant recipients, however, not all the pages of the A-133 single audit contain "findings." The GAP uses a neural network model to determine which pages in the A-133 single audit report are findings pages, which helps to identify the most important pages of that report for OIG staff.

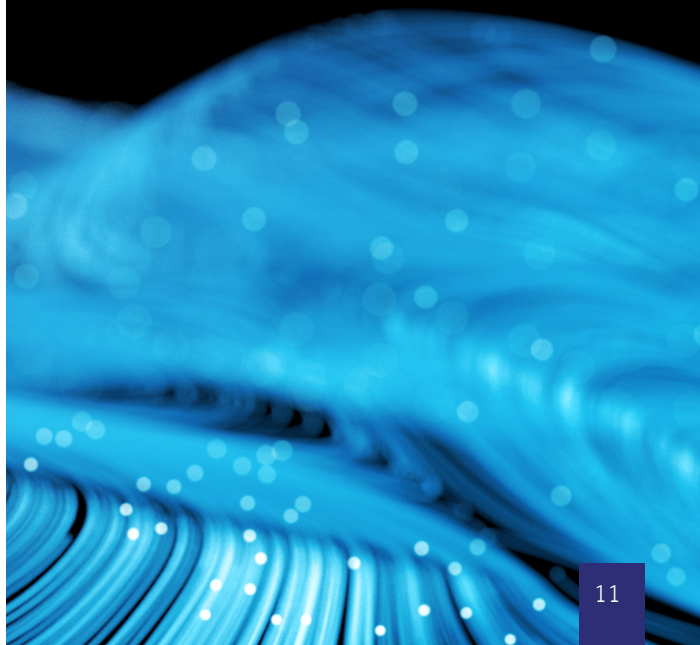
Once the findings pages are identified, text analytics techniques identify findings that aren't required to be reported on the structured data forms. Those findings are primarily financial findings about an organization's internal controls, which are very useful for OIG staff to better understand a grant recipient organization.

More Advanced Neural Network Models in the GAP (In the testing phase)

The GAP team is able to slice out specific text in an A-133 single audit finding by predicting the first and last line of audits. These models are complex as many of the reports have small changes in formatting of the text, which makes predicting the first and last line of a discovery particularly difficult. These neural network models are more advanced than the models previously deployed in the GAP. This series of models learn key markers of inconsistencies so that they can determine which text belongs to an audit finding across multiple pages of text and slice out that text accurately.

Once OIG has access to the text of individual findings, they will be able to perform more advanced text analytics such as topic modeling and clustering to classify the text and identify theme discoveries related to particular programs.

This AI solution also needs to be explainable because the government is using the outputs of these models to make decisions on whether to pursue an OIG audit or investigation. An investigation can result in civil or criminal legal action and evidence of the process used may be required. A 'black box' answer will not suffice in a court of law.



Want to know more?

If you are looking for a more technical view of XAI, check out our Data Science Lead Henry Jia's blog where he takes a deeper dive into what XAI is and how it's used.

If you would like to learn more about XAI, reach out to us.

Talk to an Expert

Meet the Authors

Claire Walsh
Director of
Engineering + Services

Henry Jia
Data Science Lead

Excella

AGILE | DATA | DIGITAL | MODERNIZATION

Further resources

Another resource for all things technical is Excellalabs.com. This is our center for all things innovation. Excella Labs is the home for Excella's technologists to share the new and experimental projects they're working on and to showcase their expertise. Take a look to explore projects, talks, and more.

About Excella

Excella is an Agile technology firm helping Washington, DC's leading organizations realize their future through the power of technology. We work collaboratively to solve our clients' biggest challenges and evolve their thinking to help them prepare for tomorrow. Together we transform bold ideas into elegant technology solutions to create real progress.
Learn more at www.excella.com.