pwc.com/me/xai

Explainable AI

Driving business value through greater understanding

#IntelligentDigital

pwc

Artificial intelligence (AI) is a transformational $15 trillion opportunity. Yet, as AI becomes more sophisticated, more and more decision making is being performed by an algorithmic ‘black box’. To have confidence in the outcomes, cement stakeholder trust and ultimately capitalise on the opportunities, it may be necessary to know the rationale of how the algorithm arrived at its recommendation or decision – ‘Explainable AI’. Yet opening up the black box is difficult and may not always be essential. So, when should you lift the lid, and how?

Contents

The $15 trillion question:

Can you trust your AI?

AI is growing in sophistication, complexity and autonomy. This opens up transformational opportunities for business and society. At the same time, it makes explainability ever more critical.

The executive view of AI on trust

67%

of the businesses leaders taking part in PwC's 2017 Global CEO Survey believe that AI and automation will impact negatively on stakeholder trust levels in their industry in the next five years.

Source: PwC 20th Annual CEO Survey, 2017

AI has entered the business mainstream, opening up opportunities to boost productivity, innovation and fundamentally transform operating models. As AI grows in sophistication, complexity and autonomy, it opens up transformational opportunities for business and society. More than 70% of the executives taking part in a 2017 PwC study believe that AI will impact every facet of business. Overall, PwC estimates that AI will drive global gross domestic product (GDP) gains of $15.7 trillion by 2030.

As businesses adoption of AI becomes mainstream, stakeholders are increasingly asking what does AI mean for me, how can we harness the potential and what are the risks? Cutting across these considerations is the question of trust and how to earn trust from a diverse group of stakeholders – customers, employees, regulators and wider society. There have been a number of AI winters over the last 30 years which have predominantly been caused by an inability of technology to deliver against the hype. However with technology now living up to the promise, the question may be whether we face another AI winter due to technologists' focus on building ever more powerful tools without thinking about how to earn the trust of our wider society.

This leads to an interesting question – does AI need to be explainable (or at least understandable) before it can become truly mainstream, and if it does, what does explainability mean?

In this Whitepaper we look at explainability for the fastest growing branch of real-world AI, that of Machine Learning. What becomes clear is that the criticality of the use case drives the desire, and therefore the need, for explainability. For example, the majority of users of recommender systems will trust the outcome without feeling the need to lift the lid of the black box. This is because the underlying approach to producing recommendations is easy to understand – 'you might like this if you watched that' and the impact of a wrong recommendation is low (a few £ spent on a bad film or a wasted 30 minutes watching a programme on catch up). However as the complexity and impact increases, that implicit trust quickly diminishes. How many people would trust an AI algorithm giving a diagnosis rather than a doctor without having some form of clarity over how the algorithm came up with the conclusion? Although the AI diagnosis may be more accurate, a lack of explainability may lead to a lack of trust. Over time, this acceptance may come from general adoption of such technology leading to a pool of evidence that the technology was better than a human, but until that is the case, algorithmic explainability is more than likely required.

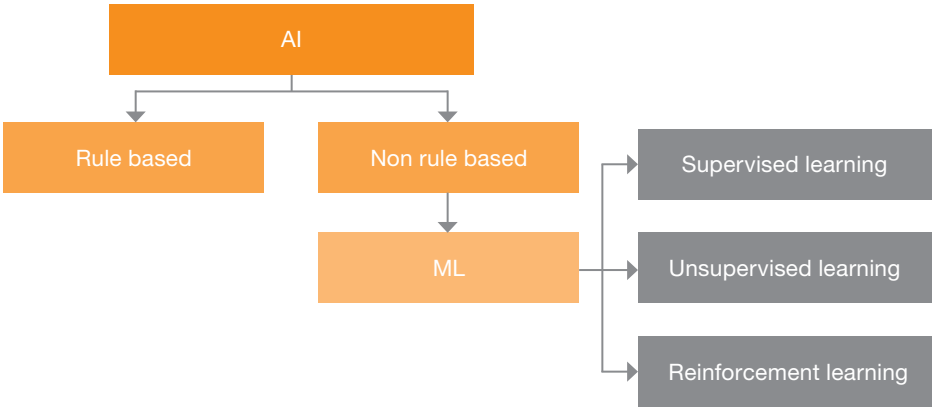
Emerging frontier

The emerging frontier of AI is machine learning (ML). For the purposes of this paper, we define Machine Learning' as a class of learning algorithms exemplified by Artificial Neural Networks, Decision Trees, Support Vector Machines, etc.: algorithms that can learn from examples (instances) and can improve their performance with more data over time. Through machine learning, a variety of 'unstructured' data forms including images, spoken language, and the internet (human and corporate 'digital exhaust') are being used to inform medical diagnoses, create recommender systems, make investment decisions and help driverless cars see stop signs We primarily focus on machine learning, a particular class of AI algorithm, because:

i) ML is the responsible for the majority of recent advances and renewed interest in AI, and

ii) ML is a statistical approach to AI that by its very nature can be difficult to interpret and validate.

Exhibit 1 | Classifying AI algorithms



Source: PwC

Operating in the dark

The central challenge is that many of the AI applications using ML operate within black boxes, offering little if any discernible insight into how they reach their outcomes. For relatively benign, high volume, decision making applications such as an online retail recommender system, an opaque, yet accurate algorithm is the commercially optimal approach. This is echoed across the majority of current enterprise AI which is primarily concerned with showing adverts, products, social media posts and search results to the right people at the right time. The 'why' doesn't matter, as long as revenue is optimised.

This has driven an approach where accuracy, above all else, has been the main objective in machine learning applications. The dominant users and researchers (often in different parts of the same large technology firms) have been concerned with the development of ever more powerful models to optimise current profits, and pave the way for future revenue streams such as self-driving cars.

In conversations with clients, we often refer to this approach (perhaps unfairly!) as 'machine learning as a Kaggle competition', referencing the popular website[1] where teams compete to build the most accurate machine learning models. In our view, this is a one dimensional vision of machine learning applications, where the biggest, latest, most complex methods vie for supremacy on the basis of a simple mathematical metric.

But what if the computer says 'No'? The absurdity of inexplicable black box decision making is lampooned in the famous (in the UK at least) 'Computer says No' sketch[2]. It is funny for a number of reasons, not least that a computer should hold such sway over such an important decision and not in any way be held to account. There is no way of knowing if it's an error or a reasonable decision. Whilst we have become accustomed to (non-AI) algorithmic decisions being made about us, despite the potential for unfairness, the use of AI for 'big ticket' risk decisions in the finance sector, diagnostic decisions in healthcare and safety critical systems in autonomous vehicles have brought this issue into sharp relief. With so much at stake, decision taking AI needs to be able to explain itself.

[1] https://www.kaggle.com/

[2] https://en.wikipedia.org/wiki/Little_Britain

Building trust

If capitalising on the $15 trillion AI opportunity depends on understanding and trust, what are the key priorities?

Explainable AI (or 'XAI') is a machine learning application that is interpretable enough that it affords humans a degree of qualitative, functional understanding, or what has been called 'human style interpretations'. This understanding can be global allowing the user to understand how the input features (the term used in the ML community for 'variables') affect the model's output with regard to the whole population of training examples. Or it can be local in which case it explains a specific decision.

Explainable AI looks at why a decision was made so AI models can be more interpretable for human users and enable them to understand why the system arrived at a specific decision or performed a specific action. XAI helps bring transparency to AI, potentially making it possible to open up the black box and reveal the full decision making process in a way that is easily comprehensible to humans.

Different groups have varying perspectives and demands on the level of interpretability required for AI. Executives are responsible for deciding the minimum set of assurances that need to be in place to establish best practices and will demand an appropriate 'shield' against unintended consequences and reputational damage. Management require interpretability to gain comfort and build confidence that they should deploy the system. Developers will therefore need AI systems to be explainable to get approval to move into production. Users (staff and consumers) want confidence that the AI system is accurately making (or informing) the right decisions. Society wants to know that the system is operating in line with basic ethical principles in areas such as the avoidance of manipulation and bias.

Organisations are facing growing pressure from customers and regulators to make sure their AI technology aligns with ethical norms, and operates within publicly acceptable boundaries.

A particular source of concern is the use of models that exhibit unintentional demographic bias. The use of explainable models is one way of checking for bias and decision making that doesn't violate ethical norms or business strategy.

Organisations have a duty to ensure they design AI that works and is robust. Adapting AI systems to fall in line with a responsible technology approach will be an ongoing challenge. PwC is helping organisations consider the ethics, morality, and societal implications of AI through Responsible AI (PwC 2017).

Benefits of interpretability

There are significant business benefits of building interpretability into AI systems. As well as helping address pressures such as regulation, and adopt good practices around accountability and ethics, there are significant benefits to be gained from being on the front foot and investing in explainability today. These include building trust – using explainable AI systems provides greater visibility over unknown vulnerabilities and flaws and can assure stakeholders that the system is operating as desired. XAI can also help to improve performance – understanding why and how your model works enables you to fine tune and optimise the model. How could better insights into business drivers such as revenue, cost, customer behaviour and employee turnover, extracted from decision making AI, improve your strategy? Further benefits include enhanced control – understanding more about system behaviour provides greater visibility over unknown vulnerabilities and flaws. The ability to rapidly identify and correct mistakes in low criticality situations add up if applied across all intelligent automaton.

Use case criticality

How far does your business need to go? When AI is used to target consumers through advertising, make investment decisions or drive cars, the required levels of interpretability will clearly vary. We believe that there are three key factors to consider when determining where interpretability is required and to what level (see Exhibit 2).

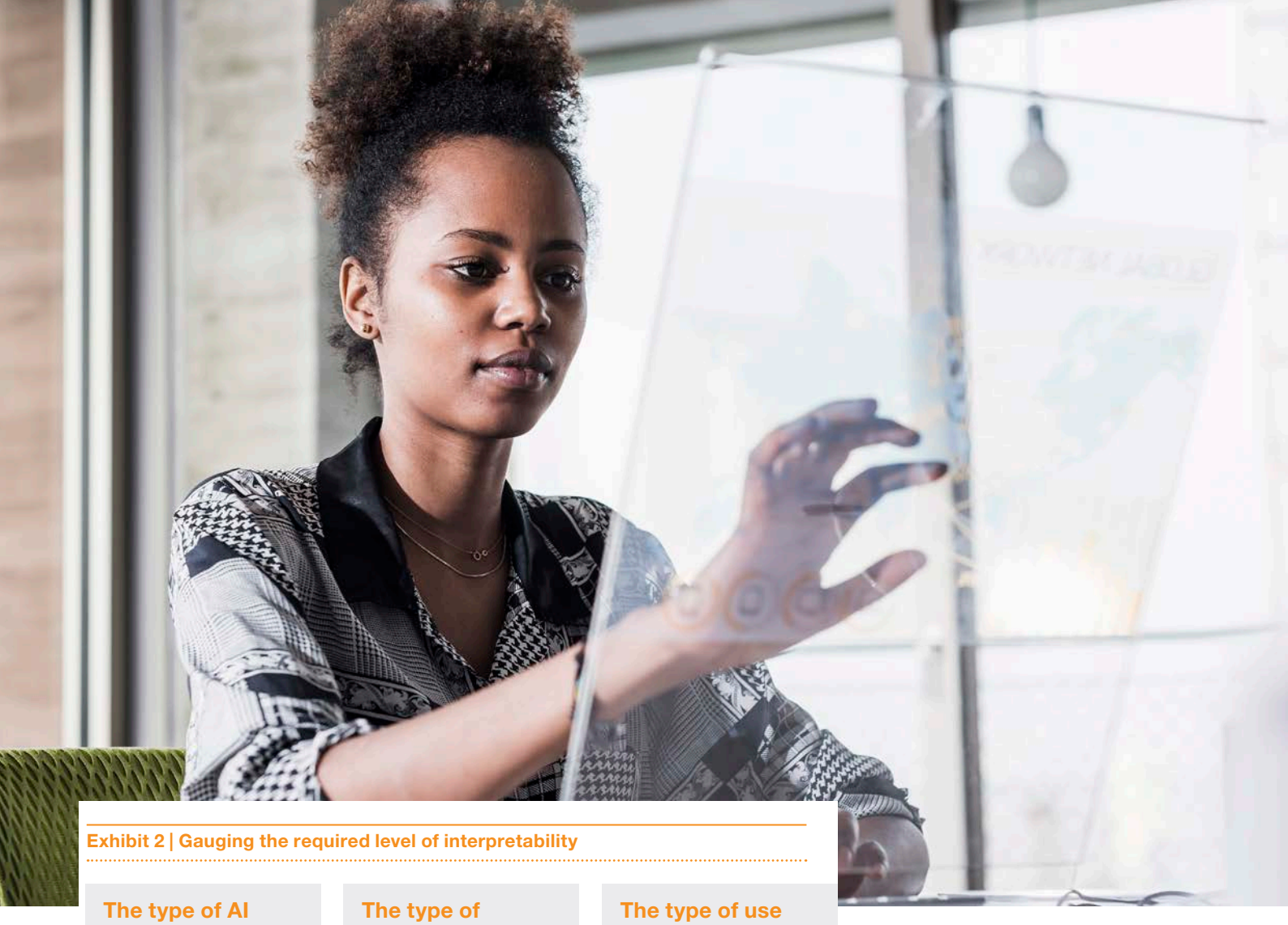


Exhibit 2 | Gauging the required level of interpretability

The type of AI

The type of AI used is the major determining factor in what level of interpretability is feasible and which techniques can be deployed. The major distinction is between rules and non-rules based systems. Non-rules based models include ML algorithms which fall into three broad categories (see Exhibit 1).

The type of interpretability

There are two dimensions to interpretability. Transparency helps shed light on black box models, whilst explainability helps organisations to rationalise and understand AI decision making:

- Explainability 'why did it do that?'
- Transparency 'how does it work?'

The type of use

The degree to which an organisation is required to be able to explain how the system works depends on the nature of the use case. There are six key domains for evaluating the criticality of use cases determined by three factors:

- Use case considerations;
- Enterprise considerations;
- Environmental considerations.

Source: PwC

The more critical the use case, the more interpretability will be required. However, the need to get inside the black box may limit the scope of the AI system – how can you balance trade-offs in areas such as increasing accuracy and performance while improving interpretability?

A thorough assessment of the utility and risk of a use case informs a set of recommendations around interpretability and risk management, which helps drive executive decision making to optimise performance and return on investment.

The way forward on interpretability

In this paper, we outline why interpretability is now vital to capitalising on AI, the key considerations for judging how explainable your AI model must be, and the business benefits from making explainability a priority. PwC's use case criticality framework helps address risks associated with a given use case and our assessment recommends optimal outcomes for interpretability, validation and verification, rigour and risk management (including bespoke controls and governance structure) tailored to your organisation. As with Responsible AI, the objective of XAI isn't to stifle or slow down innovation, but rather to accelerate it by giving your business the assurance and platform for execution you need to capitalise on the full potential of AI.

Use case criticality:

Gauging the need for explanation

The importance of explainability doesn't just depend on the degree of functional opacity caused by the complexity of your machine learning models, but also the impact of the decisions they make.

Not all AI is built equal, so it's important to think about why and when explanations are useful. Requiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable, but less capable and versatile outcomes.

For each AI use case, the verification and validation process may require different elements of interpretability, depending on the level of rigour required. Beyond the risk, it's important to consider commercial sensitivities. Opening the AI black box would make it comprehensible for users, but it could also give away valuable intellectual property.

Exhibit 3 | The need for Explainability

Working as intended?

A primary driver for model interpretability is the need to understand how a given model makes predictions, while ensuring that it does so according to the desired specifications and requirements demanded of it.

How sensitive is the impact?

The need for interpretability also depends on the impact. For an AI system for targeted advertising, for example, a relatively low level of interpretability could suffice, as the consequences of it going wrong are negligible. On the other hand, the interpretability for an AI based diagnosis system would be significantly higher. Any errors could not only harm the patient, but also deter adoption of such systems.

Are you comfortable with the level of control?

The other key piece in the jigsaw is the level of autonomy. Does the AI system make decisions and perform actions consequently, or do its outputs function as mere recommendations to human users who can then decide whether or not to follow these? Explainable factors can be used for a level of rules based control or to flag to humans automatically. It's important to be able to fully understand a system before allowing it to make business decisions without human intervention.

Use case criticality

All these various considerations come together to determine whether explainability is necessary and the level of rigour that would need to be applied.

Source: PwC

Exhibit 4 outlines the main use case criticality evaluation criteria across six domains. In practice, the criticality of use case explainability is driven predominantly by three economic factors:

1. The potential economic impact of a single prediction;
2. The economic utility of understanding why the prediction was made with respect to the choice of actions that may be taken as the result of the prediction;
3. The economic utility of the information gleaned from understanding trends and patterns across multiple predictions (management information).

However, organisations must place a higher value on the importance of factors beyond the economic and technical drivers such as executive risk, reputation, and rigour.

Exhibit 4 | Use case criticality components

Revenue

The total of the economic impact of a single prediction, the economic utility of understanding why a single prediction was made, and the intelligence derived from a global understanding of the process being modeled.

Rate

The number of decisions that an AI application has to make e.g. two billion per day versus three per month.

Rigour

The robustness for the application: its accuracy and ability to generalise well to unseen data.

Regulation

The regulation determining the acceptable use and level of functional validation needed for a given AI application.

Reputation

How the AI application interacts with the business, stakeholders, and society and the extent a given use case could impact business reputation.

Risk

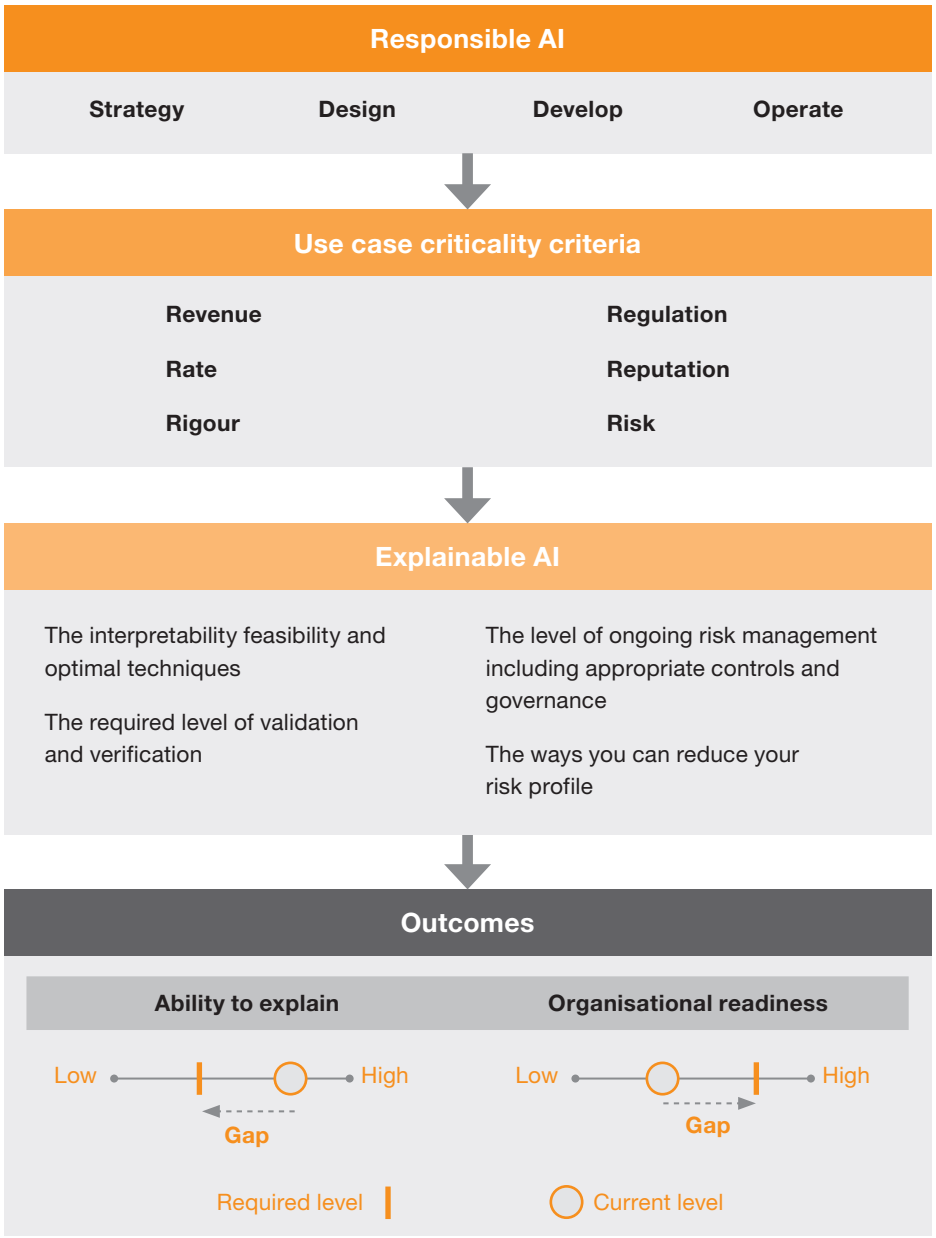
The potential harm due to an adverse outcome resulting from the use of the algorithm that goes beyond the immediate consequences and includes the organisational environment: executive, operational, technology, societal (including customers), ethical, and workforce.

Source: PwC

Exhibit 5 outlines PwC's approach to assessing the criticality of an AI use case. Explainable AI works in conjunction with PwC's overarching framework for the best practice of AI: Responsible AI, which helps organisations deliver on AI in a responsible manner. Upon evaluation of the six main criteria of use case criticality, the framework will recommend for new use cases the optimal set of recommendations at each step of the Responsible AI journey to inform Explainable AI best practice for a given use case. Whilst for existing AI implementations, the outcome of the assessment is a gap analysis showing an organisation's ability to explain model predictions with the required level of detail compared to PwC leading practice and the readiness of an organisation to deliver on AI (PwC Responsible AI, 2017).

The gap analysis provides a view that informs the trade-off between prediction accuracy and explainability. If an organisation is ahead of the required level for the ability to explain, this would suggest the organisation has room to trade-off some level of additional explainability for increased model accuracy. In the same context, a scenario where the ability to explain falls below the required level would result in a need to reduce the model prediction accuracy in order to achieve greater explainability.

Exhibit 5 | Use case criticality



Source: PwC

Key considerations for explainability:

How to embark on explainability

Given a compelling rationale for incorporating explainability into an application, how do you choose the appropriate machine learning algorithm, explanation technique, and method for presenting the explanation to a human?

Explainable AI makes it possible to open up the black box and reveal the aspects of the decision making process that provide a meaningful explanation to humans. This however comes with the need for additional software components and application design considerations.

Explainable by design

As with most engineering processes, you must consider the capabilities your system requires at the early stages of the design phase. Explainability is no different – it needs to be considered up front and embedded into the design of the AI application. It affects the choice of machine learning algorithm and may impact the way you choose to pre-process data. Often this comes down to a series of design trade-offs.

Trade-off between performance and explainability

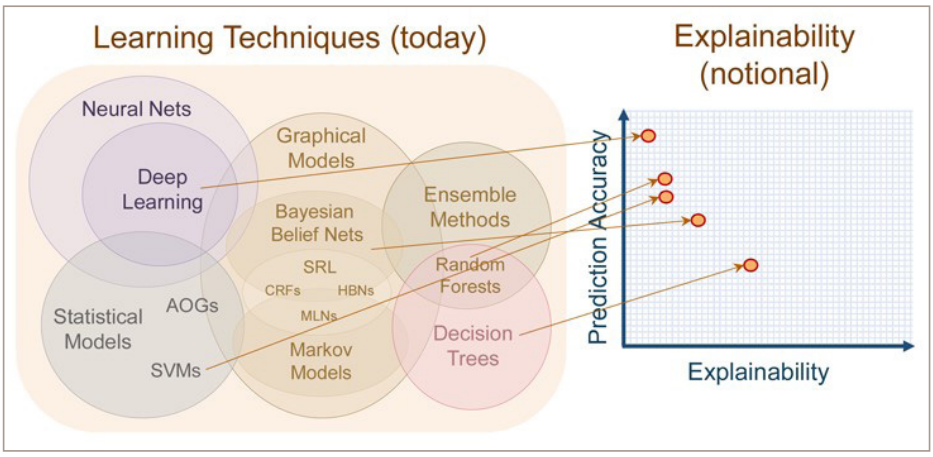
Explainable systems usually incorporate some sort of model interpreter. We can think of interpretation as a method for mapping of a concept (e.g. 'cat') to input features that a human can make sense of (e.g. group of pixels representing whiskers). The explanation is the collection of interpretable features that contribute to the decision (e.g. whiskers + tail + collar ⇒ cat). Thus, explainability is linked directly to model interpretability – the degree to which the interpreter can assign interpretable features to a model prediction.

Interpretability is a characteristic of a model that is generally considered to come at a cost. As a rule of thumb, the more complex the model, the more accurate it is, but the less interpretable it is. See Exhibit 6: Relative explainability of learning algorithms. Complexity is primarily driven by the class of machine learning algorithm used to generate a model (e.g. Deep Neural Network vs Decision Tree) as well as its size (e.g the number of hidden layers in a neural network).

Some models however, such as decision trees, are highly amenable to explanation. It is possible to build commercially useful models where the entire decision process can be diagrammatically illustrated. If the model becomes too large for useful graphical representation, the tree structure of the model means that interpreter software can trace clear decision paths through the model and extract the key determinants of a prediction. Neural networks on the other hand, whilst amenable to graph analysis, contain many more connections and have more subtle properties with respect to node interactions that are inherently difficult to interpret.

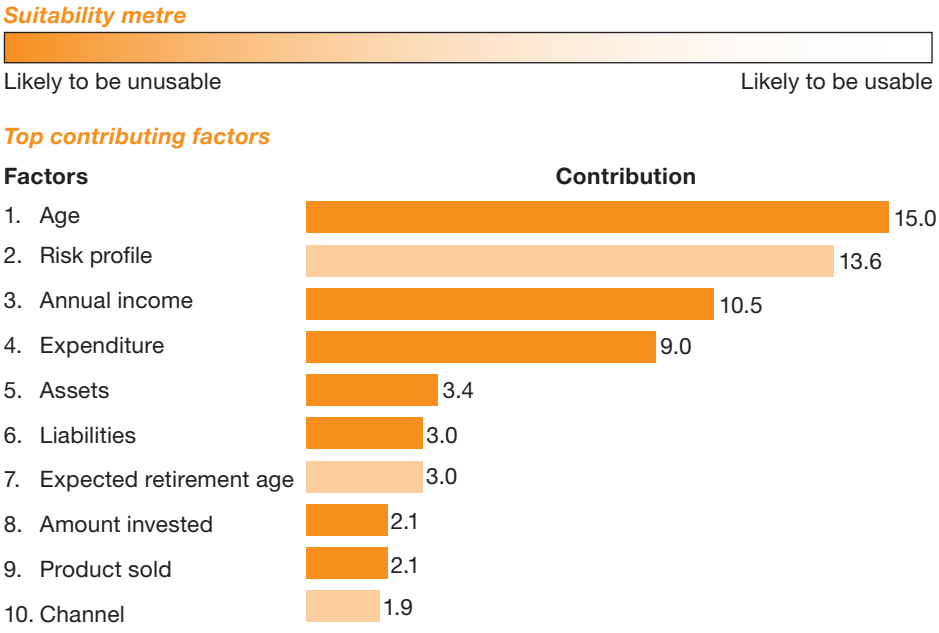
It's worth noting that some researchers are challenging this long held view (such as Montavon et al. 2017), who argue that recent advances in interpreting neural networks allow users deeper insights unavailable from simple models because of their complexity. This makes intuitive sense when, as we shall see later, with features that comprise of Deep Neural Nets (DNNs), explanations can include representations of complex concepts as images that can't be meaningfully represented by a simple linear model.

Exhibit 6 | Relative explainability of learning algorithms



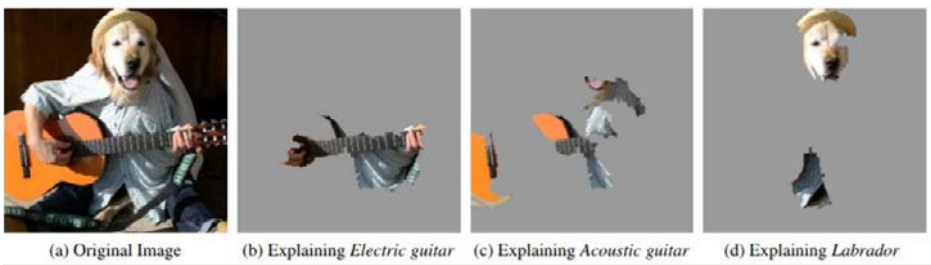
Source: DARPA

Exhibit 7 | Feature importances in investment product suitability



Source: PwC

Exhibit 8 | Pixel importance in explaining image recognition



Source: Tulio Ribeiro et al (2016)

Exhibit 9 | Text importance to explain article saliency

Source: PwC, Original article: Bloomberg

Interpretability is a necessary but not sufficient condition for explanation. A model interpreter may generate representation of a decision process, but turning this into a useful explanation can be challenging for the following reasons:

The limits of explainability

The type of learning algorithm that generates a model is a key factor in determining explainability. But the type of explanation required, and the type of input data used in the model can be equally important.

Feature importances

Most explanations are limited to a list or graphical representation of the main features that influenced a decision and their relative importance. Exhibits 7, 8 and 9 show typical approaches to explanation presentation.

Presenting feature importances in these ways largely ignores the details of interactions between features, so even the richest explanations based on this approach are limited to relatively simple statements. Domain specific logic can be applied to explanations and presented as text with a natural language generation (NLG) approach. PwC has successfully implemented this method in domains such as corporate credit rating in which rich textual explanations are automatically generated. Here is an excerpt from an automatically generated credit report: 'Given the company's financials, media references linking it with acquisitions is associated with credit strength'.

As you can see, being automatically generated, the grammar is not perfect, however statements such as these can be aggregated and more domain knowledge applied to build more ever more complex explanations.

Problem domain

Certain types of problem can't be readily understood by quantifying a handful of factors. For example in the field of genomics, the prediction is usually almost entirely a driven by the combinations' nucleotides. Here, simple feature importance conveys little by way of an explanation, although richer explanatory representations can be informative (see Vidovic et al 2016).

Data preprocessing

Many ML applications employ significant data pre-processing to improve accuracy. For example, dimensionality reduction through principal component analysis (PCA), or using word vector models on textual data can obscure the original human meanings of the data making explanations less informative.

Correlated input features

When highly correlated training data is available, correlates are routinely dropped during the feature selection process or simply swamped by the other correlates with more predictive power. Hence ambiguity concerning the latent factor behind the correlates is hidden from the explanation, potentially leading the user to an incorrect conclusion.

Type of prediction

Certain types of model prediction are easier to explain than others. Binary classifiers are the easiest (i.e. 'Is it X or is it Y?'). We can talk about factors that push the decision in one way or the other. This can be extended to ordinal multilabel classifiers (e.g. predicting credit ratings) and regressors (e.g. predicting store sales) that have a natural 'direction' in the output (magnitude of risk or revenue). This allows straightforward application of domain specific knowledge to enrich explanations. In contrast, multi-label classifiers in domains where there is little or no inherent order (as in many image classification problems), are generally limited to 'one vs all' type explanations.

Human explanations are often flawed:

It's worth noting that, to some degree, human explanations suffer from the limits described above. We are prone to oversimplification, cognitive biases, and have difficulty explaining abstract concepts unless specific language already exists. Whilst we can easily recognise the difference between many common items, we can have difficulty defining the differences (e.g. we can see a pair of twins are different, but can't quite explain why). More worryingly, research by psychologists such as Gazzaniga (2012)[3] show that human post-hoc explanations of their own behaviour can be grossly inaccurate, with subconscious processes creating consistent narratives to support a positive sense of self rather than reporting the 'facts'. Thus, even the best intentioned human explanations can be completely unreliable unbeknownst to the explainer.

What explanation techniques are available?

A number of methods for generating explanations exist, ranging from classic black box analysis approaches that have been used in science and engineering for generations, to the latest methods designed for Deep Neural Networks (DNNs). We cannot list these exhaustively, but provide a summary of the more popular approaches we see used by the machine learning community, as well as some promising recent developments.

We can broadly group explanation techniques in model-agnostic approaches and learning algorithm specific approaches. Model-agnostic approaches are essentially 'black box' explainers and can in principle be applied to any ML model. They do not need to see what is going on under the hood, just tweak the inputs and observe the effects. This can come at the cost of less explainability than algorithm specific techniques that more directly probe a model's inner workings. Nonetheless, these approaches are often very effective and straightforward to implement.

Sensitivity analysis:

This is an approach that is applied in many domains to understand the behaviour of not just models, but any opaque, complex system, such as electrical circuits. Whilst there are a number of formal approaches that are designed to give particular statistical insights, the simplest approach is to marginally alter (perturb) a single input feature and measure the change in model output. This gives a local, feature specific, linear approximation of the model's response. By repeating this process for many values, a more extensive picture of model behaviour can be built up. This approach is often extended to Partial Dependence Plots (PDP) or Individual Conditional Expectation (ICE) Plots to give global graphical representation of single feature importances. Extending basic sensitivity analysis into higher dimensions is trivial: multiple rather than a single features are perturbed to build a composite picture of feature importances.

[3] Who's in Charge?: Free Will and the Science of the Brain

The benefits of this method are its simplicity and ease of implementation. In many problem domains, the results are very intuitive. It works particularly well for simple models with smoothly varying behaviour and well separated features. However, its simplicity means that it can be applied to complex models such as DNNs where it is effective for extracting pixel importance in image recognition explanations.

It's important to note that sensitivity analysis gives explanations for the variation in the model output rather than the absolute value. Usually this provides sufficient explanation: the question 'what makes this image more cat-like?' is essentially indistinguishable from 'what makes this image a cat?'.

This approach has several drawbacks: it doesn't directly capture interactions amongst features, and simple sensitivity measures can be too approximate. This can be potentially problematic for discontinuous features such as categorical information and one hot encoding frequently used in natural language processing.

Local Interpretable Model – Agnostic Explanations – LIME[4]

LIME addresses the main shortcomings of basic sensitivity analysis. Like sensitivity analysis, it can be applied to any model, but unlike sensitivity analysis it captures feature interactions. It does so by performing various multi-feature perturbations around a particular prediction and measuring the results. It then fits a surrogate (linear) model to these results from which it gets feature importances, capturing local feature interactions. It can also handle non-continuous input features frequently found in machine learning applications. An open source implementation of this approach is available[5] which currently makes this the 'go to' interpreter for many practitioners.

Shapley Additive Explanations – SHAP[6]

Similarly to LIME, SHAP is a local surrogate model approach to establishing feature importance. It uses the game theoretic concept of Shapley values to optimally assign feature importances. The Shapley value of a feature's importance is its average expected marginal contribution after all possible feature combinations have been considered.

The Shapley value guarantees to perfectly distribute the marginal effect of a given feature across the feature values of the instance. Thus SHAP currently produces the best possible feature importance type explanation possible with a model agnostic approach. This however comes at a cost: the computational requirements of exploring all possible feature combinations grow exponentially with the number of input features. For the vast majority of problems, this makes complete implementation of this approach impractical and approximations must suffice.

The methods outlined above can be applied to any class of model, however, richer and more accurate explanations are often available with learning algorithm specific interpreters.

Tree interpreters

As discussed earlier, decision trees are a highly interpretable class of model, albeit one of the least accurate. The Random Forest algorithm is an extension of the basic decision tree algorithm which can achieve high accuracy. It is an ensemble method that trains many similar variations of decision trees and makes decisions based on the majority vote of individual trees. A decision tree interpreter can be applied to individual trees in the forest and the feature importances aggregated. Thus, random forests are highly interpretable models with high accuracy that can be understood both globally and locally.

This makes random forests the 'go to' algorithm in many commercial applications. A widely used open source tree interpreter package 'treeinterpreter' is available that makes this approach relatively straightforward to implement[7], although significant domain expertise may be required for multi-label classification problems.

Neural Network Interpreters

Neural networks, particularly DNNs, are characterised by their complexity and consequently are widely considered difficult to interpret. In our opinion, the issue is more one of a lack of generalisability. There is no inherent barrier to gaining insight into the inner workings of neural networks, but it often needs to be tackled on a case by case basis requiring significantly more expertise and effort than say the relatively trivial exercise of extracting global feature importances from a random forest.

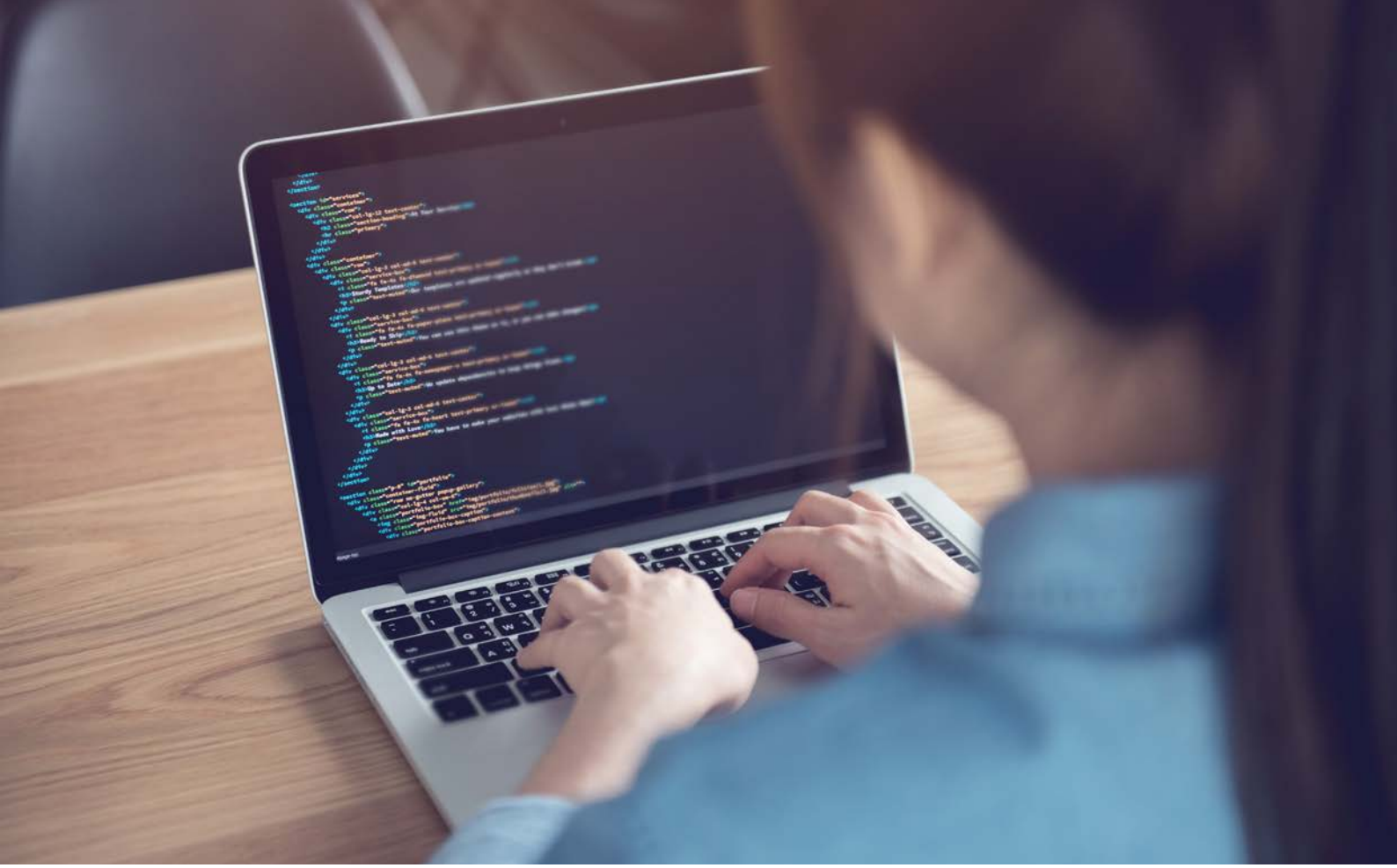
This effort is often rewarded with sometimes surprising and informative insights into how the DNN decomposes a problem into hierarchies (e.g. fine texture vs. gross structure in image classifiers), manifolds (simplified representations of more complex concepts) and class label 'prototypes' (idealised representations of target classes).

[4] Tulio Riberio et al. (2016)

[5] http://github.com/marcotcr/lime

[6] Lundberg and Lee (2017)

[7] http://blog.datadive.net/random-forest-interpretation-with-scikit-learn/



Activation maximisation (AM)

Is a method that can be used to find a DNNs prototypical representation of a particular concept. For example, in an image recognition DNN, AM could be used to produce a visual representation of the DNNs concept of a cat. By searching for input patterns that maximise a particular output (e.g. the probability the image is a cat), a prototypical cat image can be extracted. Thus, AM can afford direct interpretable insight into the internal representations of DNNs.

Whilst sensitivity analysis is commonly used to explain feature importance in DNNs, a neural network specific approach called **Relevance Propagation** can produce more accurate, robust explanations. It can be thought of the inverse of sensitivity analysis in that the algorithm starts at the model output, and works back through the network, assigning relevance of inputs from the preceding layers (when fed-forward) until it reaches the input layer. Whilst there are many variations on this approach, they tend to be implemented on a bespoke basis. However, an open source implementation called DeepLIFT[8] is available for those who would rather not build an interpreter from scratch.

In the last section, the trade-offs, limits and methods for explainable AI were discussed. In the next section we extend the discussion to the implications for model evaluation – this is key to the safe and effective use of AI. In our opinion, even slight improvements in the model evaluation process can pay dividends in future model performance and reducing the risk of adverse model behaviour.

Model evaluation: going beyond statistical measures

Machine Learning model evaluation is critical to validate that systems meet the intended purpose and functional requirements. The de facto approach amongst ML practitioners has been to test models on a held-out portion of the training data and report error. Graphical analysis of confusion matrices, ROC curves and learning curves can further enhance the tester's understanding of the model's behaviour.

Beyond these quantitative approaches, a functional understanding of ML model behaviour using XAI can give critical insight not available through quantitative validation approaches. An example is a study carried out in the nineties using rules based learning and neural networks to decide which pneumonia cases should be admitted to hospital or treated at home. The models were trained on patients' recovery in historical cases[9].

[8] Shrikumar et al (2017). See also: https://github.com/kundajelab/deeplift

[9] Caruana, et al. (2015)

Both models predicted patient recovery with high accuracy with the neural network found to be the most accurate. Both models predicted that pneumonia patients with asthma shouldn't be admitted because they had a lower risk of dying.

In fact, pneumonia patients were at such high risk, they were routinely admitted directly to the intensive care unit, treated aggressively, and as a consequence had a high survival rate. Because the rules based model was interpretable, it was possible to see that the model had learnt 'if the patient has asthma, they are at lower risk'. Despite ostensibly good quantitative performance metrics, the counterintuitive explanation invalidated the model from a clinical point of view. Had this model been deployed in production, it could have led to unnecessary patient risk and possibly death.

Thus a functional explanation, even if a gross approximation of the underlying model complexity, can catch the type of potentially dangerous informational shortcuts machine learning algorithms are good at finding in contravention of the developers intentions.

Whilst the previous example is unlikely to occur these days, there are many examples of badly trained algorithms using data such as people's names to infer demographic characteristics, using URLs in mined text to classify documents and using copyright tags to classify images. These models may perform well in testing, but are liable to catastrophic failure or unintended consequences in production. This type of oversight can be avoided with even a fairly rudimentary deployment of XAI.

Continuous evaluation

Unlike traditional software and hand crafted top down models, machine learning models may be periodically retrained or continuously updated (online learning) as they learn from new instances. Factoring explainable AI outputs into automated controls means robust qualitative rules based safeguards against unexpected, unwanted, and known weakness in model behaviour can be applied. For example, if a single pixel in an image is identified as the most important feature in a decision, the model could be the target of an adversarial attack by criminals. Applying the rule 'if a single pixel is the primary explanatory feature by a large margin, then raise alert' could safeguard against such attacks. Whether the reason is an adversarial attack or not, XAI has alerted us to problem, whatever the nature, through the application of common sense rules.

2.12 Model transparency

Model transparency is concerned with conveying to a user the structural details of the model, statistical and other descriptive properties of the training data, and evaluation metrics from which likely behaviour can be inferred. Irrespective of the need for explainable AI or not, this is basic information without which it is impossible to make informed use of a machine learning algorithm.

In cases where engineers are charged with making someone else's model explainable, as may be the case with a commercially available off the shelf model, the model's users should take into account the degree of transparency that comes with the model. Selecting an optimal approach to XAI is significantly more straightforward when the details of the model are well known rather than having to deal with a mystery black box.

[9] Caruana, et al. (2015)

Business benefits

Turning explainability into a competitive differentiator

How XAI can not only strengthen stakeholder confidence, but also improve performance, make better use of AI and stimulate further development.

The greater the confidence in the AI, the faster and more widely it can be deployed. Your business will also be in a stronger position to foster innovation and move ahead of your competitors in developing and adopting next generation capabilities.

Exhibit 10 | Eight business benefits of Explainable AI

Optimise	Retain	Maintain	Comply
Model performance	Control	Trust	Accountability
Decision making	Safety	Ethics	Regulation

Source: PwC

Optimise

Model performance

One of the keys to maximising performance is understanding the potential weaknesses. The better the understanding of what the models are doing and why they sometimes fail, the easier it is to improve them. DeepMind was able to improve and optimise AlphaGo (the machine that famously beat the world's best Go player) through being able to see and understand how and why the system was making particular decisions. It would not be possible to fully optimise AlphaGo and create the success we know today if the system was functioning as a black box.

Explainability is a powerful tool for detecting flaws in the model and biases in the data which builds trust for all users. It can help verifying predictions, for improving models, and for gaining new insights into the problem at hand. Detecting biases in the model or the dataset is easier when you understand what the model is doing and why it arrives at its predictions.

Decision making

As discussed earlier, the primary use of machine learning applications in business is automated decision making. However, often we want to use models primarily for analytical insights. For example, you could train a model to predict store sales across a large retail chain using data on location, opening hours, weather, time of year, products carried, outlet size etc. The model would allow you to predict sales across my stores on any given day of the year in a variety of weather conditions. However, by building an explainable model, it's possible to see what the main drivers of sales are and use this information to boost revenues.

A popular use of machine learning is predicting customer churn. A 90% accurate prediction that Joe Bloggs (your customer) will switch to a competitor in the next month is interesting. But so what? Perhaps you could offer a discount if you want to retain the customer, but what if they are switching because they had a bad customer service experience and aren't particularly price sensitive in this instance. You may unnecessarily offer them (and many others) ineffective and expensive incentives without realising you are losing customers for a reason that has nothing to do with price. If the churn model could make a 90% accurate prediction with a reason 'the customer will switch to a competitor because they spent 47 minutes waiting for customer service to answer the phone over the past 12 months,' then fixing the root cause (customer service) is by far the better strategy.

Exhibit 11 | Reinforcement learning: AlphaGo Zero

When using explainable AI systems, we can try to extract this distilled knowledge from the AI system in order to acquire new insights. One example of such knowledge transfer from AI system to human arose when DeepMind's AI model AlphaGo identified new strategies to play Go, which certainly now have also been adapted by professional human players.

What is remarkable about AlphaGo is that the model didn't require any human input at all. It wasn't given any rules; instead it mastered the game of Go by playing alone and against itself.

So how does it work? There are three main components:

1. The Policy Network (trained on high level games to imitate players);
2. The Value Network (evaluate the board position and calculate the probability of winning in a given position);
3. The Tree Search (looks through different variations of the game to try and figure out what will happen in the future).

First the Policy Network scans the position and comes up with the interesting spots to play and builds up a tree of variations. It then deploys the Value Net to tell how promising the outcome of this particular variation is with the goal of maximising the probability of winning.

The extraordinary progress of this form of reinforcement learning offers immense potential to support human decision making. Instead of playing Go or Chess, an AI model in the right environment can 'play' at corporate strategy, consumer retention, or designing a new product.

Source: Mastering the Game of Go without Human Knowledge, Silver et al. 2017

Retain

Control

To move from proof of concept to fully-fledged implementation, you need to be confident that your system satisfies certain intended requirements, and that it does not have any unwanted behaviours. If the system makes a mistake, organisations need to be able to identify that something is going wrong in order to take corrective action or even to shut down the AI system.

XAI can help your organisation retain control over AI by monitoring performance, flagging errors and providing a mechanism to turn the system off. From a data privacy point of view, XAI can help to ensure only permitted data is being used, for an agreed purpose, and make it possible to delete data if required.

Developers frequently try to solve problems by 'throwing data' at AI in instances of a black box system. Having visibility over the data and features AI models are using to provide an output can ensure that issues arise can be understood and a level of control can be maintained. For systems that learn through customer interactions, interpretable AI systems can shed light on any adverse training drift.

Safety

There have been several concerns around safety and security of AI systems, especially as they become more powerful and widespread. This can be traced back to a range of factors including deliberate unethical design, engineering oversights, hacking and the effect of the environment the AI operates in.

XAI can help to identify these kinds of faults. It's also important to work closely with cyber detection and protection teams to guard against hacking and deliberate manipulation of learning and reward systems.

Maintain

Trust

Building trust in artificial intelligence means providing proof to a wide array of stakeholders that the algorithms are making the correct decisions for the right reasons. Explainable algorithms can provide this up to a point, but even with state of the art machine learning evaluation methods and highly interpretable model architectures, the context problem persists: AI is trained on historical datasets which reflect certain implicit assumptions about the way the world works. Events can occur that radically reframe problems (an earthquake, a new central bank policy, a new technology) and make the historical training data invalid. By gaining an intuitive understanding of a model's behaviour, the individuals responsible for the model can spot when the model is likely to fail and take the appropriate action.

XAI also helps to build trust by strengthening the stability, predictability and repeatability of interpretable models. When stakeholders see a stable set of results, this helps to strengthen confidence over time. Once that faith has been established, end users will find it easier to trust other applications that they may not have seen. This is especially important in the development of AI, as models are likely to be deployed in settings where their use may alter the environment, possibly invalidating future predictions.

Ethics

In software development life cycles, engineers and developers are focused on functional requirements while business teams are focused on speeding up AI implementation and boosting performance. The ethical impact and other unintended consequences can easily be obscured by the need to meet these pressing objectives.

It's important that a moral compass is built into the AI training from the outset and AI behaviour is closely monitored thereafter through XAI evaluation. Where appropriate, a formal mechanism that aligns a company's technology design and development with its ethical values and principles and risk appetite may be necessary.

PwC is working towards embedding ethical considerations into the design phase of the AI development cycle, with clear governance and controls to guard against risks emerging from ethical oversight. Accountability must be shared between business managers and solution owners with a view of understanding and mitigating risks early on.

Comply

Accountability

It's important to be clear who is accountable for an AI system's decisions. This in turn demands a clear XAI-enabled understanding of how the system operates, how it makes decisions or recommendations, and how it learns and evolves over time and how to ensure it functions as intended.

To assign responsibility for an adverse event caused by AI, a chain of causality from the AI agent back to the person or organisation needs to be established that can be reasonably held responsible for its actions. Depending on the nature of the adverse event, responsibility will sit with different actors within the causal chain that lead to the problem. It could be the person who made the decision to deploy the AI for a task to which it was ill suited, or it could rest with the original software developers who failed to build in sufficient safety controls.

Regulation

While AI is lightly regulated at present, this is likely to change as its impact on everyday lives becomes more pervasive. Regulatory bodies and standard institutions are focusing on a number of AI-related areas, with the establishment of standards for governance, accuracy, transparency and explainability being high on the agenda. Further regulatory priorities include safeguarding potentially vulnerable consumers.

It's important for industry bodies and regulators to bring regulation into line with developments in AI and mirroring the AI ecosystem within their sectors. This of course means that many of these bodies will need to acquire the requisite expertise to effectively discharge this duty. There is an important role for professional and academic bodies representing the technology community (for example ACM and IEEE) to play as technical experts and help advise policy.

However, the current risks from AI come from its deployment in specific industrial contexts, and current regulatory systems are generally far better placed to monitor, and sanction their constituents. PwC is actively advising regulators, for instance the Financial Conduct Authority (FCA), on the impact of current developments of the field of AI including the creation of safe 'sandbox' environments for live testing. Working closely with the British Standards Institution (BSI) and the International Organisation for Standardisation (ISO), PwC is identifying requirements for specific areas of AI/ML to help develop a new set of standards for AI.

Exhibit 12 | GDPR for Machine Learning

The EU's General Data Protection Regulation (GDPR) is likely to create a host of new and complex obligations between data subjects and models, particularly around machine learning (ML). This section attempts to address some of these questions:

1. Does the GDPR prohibit machine learning?

Technically, The GDPR contains a prohibition on the use of automated decision-making without human intervention. However, this prohibition is circumvented where processing is contractually necessary, authorised by another law, or the data subject has explicitly consented.

2. Is there a 'right to explainability' from ML?

Much of the predictive power of ML lies in complexity that's difficult, if not impossible, to explain. Articles 13-15 set out a right to 'meaningful information about the logic involved', and data subjects are entitled not to be subject to (Article 22), to receive an explanation of, and to challenge decisions (Recital 71). While these provisions may establish the need for a detailed explanation of a model's workings, regulators are more likely to focus on a data subject's ability to make informed decisions.

3. Do data subjects have the ability to demand that models be retrained without their data?

If consent is withdrawn for the data used in an ML model, the model might theoretically have to be retrained on new data. However all processing that occurred before the withdrawal remains legal. If the data was legally used to create a model or prediction, training data can be deleted or modified without affecting the model. It may technically be possible to rediscover original data (see Nicolas Papernot et. al), however this is unlikely in practice, and it is not expected that models will be subject to constant demands of being re-trained on new data.

This considers just some of the complex intersections between ML and the GDPR and these questions remain extremely nuanced. Lawyers and privacy engineers are going to be a central component of data science programs in the future.

Source: Information Commissioner's Office

XAI is only going to get more important

AI is advancing, deployment is proliferating and the focus of regulators, customers and other stakeholders is increasing in step. XAI can help you keep pace.

We are moving out of the carefree days of silicon valley giants getting content in front of eyeballs and into a broad, industrial revolution where things are about to get a lot more serious…. Any cognitive system allowed to take actions on the back of its predictions had better be able to explain itself, if even in a grossly simplified way.

Explainability has been largely ignored by the business community and is something PwC is helping organisations to solve. There are no perfect methods, and some problems are inherently not semantically understandable, but most business problems are amenable to some degree of explanation that de-risks AI, builds trust, and reduces overall enterprise risk. Making XAI a core competency and part of your approach to AI design and QA will pay dividends today and in the future.

PwC is helping advise and assure our clients by building and helping organisations build AI systems that are explainable, transparent, and interpretable and by assessing and assuring organisations have built AI modes that adhere to our standards.

Key takeaways:

1. AI must be driven by the business

Developers are mostly focussed on delivering well-defined functional requirements, and Business Managers on business metrics and regulatory compliance. Concerns around algorithmic impact tend only to get attention when algorithms fail or have a negative impact on the bottom line. Because AI software is inherently more adaptive than traditional decision-making algorithms, problems can unfold with quicker and greater impact. Explainable AI can forge the link between non-technical executives and developers, allowing the effective transmission of top level strategy to junior data scientists. Insufficient governance and quality assurance around this technology is inherently unethical and needs to be addressed at all levels of the organisation. Without XAI, governance is very difficult.

2. Executive accountability

With the proliferation of AI systems and, equally, the increased impact on organisations, individuals and society, it needs to be clear who is accountable for an AI system's decisions. If executives are required to accept accountability for AI, they will need to understand the risk it introduces to their business. Without a deeper understanding of the system's rationales, executives would introduce unknown risks to their risk profile. In order to accept accountability, executives must have the confidence that a system operates within defined boundaries.

3. Doing the right thing, right

Humans are the designers of AI systems, and ethics are embedded before the very first line of code is written. Hence it is imperative to have defined ethics and core values, along with a governance system that ensures compliance, before the development and deployment of an AI system. These foundations help guide your organisation when engaging with potential customers, restricting them from unethical misuse. It's important to ensure that business managers understand, are accountable for the risks and, where appropriate, a formal mechanism is in place that aligns your technology with its ethical policies and risk appetite.

4. Future-proofing your AI

The need to be interpretable is increasing. In sectors such as financial services, the use of advanced AI is already so well entrenched that risks should be at the top of the risk register. Other sectors such as healthcare and transport are fast following suit. And, as AI continues to permeate through the economy, all sectors would eventually need to judge the criticality and impact of their AI on the one side and how much faith they have in the outcomes on the other. A new wave of AI specific regulation is also coming. In this respect, AI explainability is now coming to rank alongside cyber as a threat, but also a valuable differentiator if handled smartly.

5. Explainable by design

You might assume that this isn't an immediate priority as the systems you're using are still fairly rudimentary and humans still have the last word on key decisions. But how long will this be the case and how can you lay secure foundations for moving ahead? The inability to see inside the black box can only hold up AI development and adoption. You might assume that the necessary layers of understanding and control can be applied as your systems develop. But that would simply leave you with the same jumble of bolt-ons that have made most technology infrastructures so difficult to manage and optimise. It's important to begin thinking about interpretability now as AI becomes more prevalent and complex. This is akin to AI safety research, investing the time and effort now to be prepared for the future. And by proactively putting in place the right measures early, you can future-proof AI assurance rather than relying on decades of reactive fixes.

In conclusion being able to explain not just how the black box works, but your approach to designing, implementing and running the black box will increasingly be a key requirement of driving trust in AI. This is where the worlds of Responsible AI and Explainable AI collide. However, it is likely that you will have two classes of AI, those solutions that are already active and those that you will deploy in the future. For those active applications you should look to understand whether your ability to explain the results aligns with the expectations of the key stakeholders and whether you need to make any changes. It is important to identify any gap in understanding at both an algorithm and a process level. For those applications that are either on the drawing board or are yet to be identified, you need to look at how you will ensure you have a mechanism for understanding the level of explainability required for an algorithm in the idea and design phase for new AI applications.

Explainability is a business issue that needs a business and board response, and not just a tech response. The time to act is now – get this right, and you can move forward with greater confidence.

Appendix 1

Classifying machine learning algorithms

Exhibit 1 | Classifying machine learning algorithms

Supervised learning

The most common approach to machine learning where the goal is to train a classifier or regressor by finding function that maps the training examples to training label with minimal error. Supervised learning is used in situations where the training data examples are labelled (e.g. image of a cat with label 'cat') and the instances encountered in production expected to be drawn from a similar distribution of instances used in training. There are many applications ranging from image recognition, to spam detection, to stock price prediction.

Unsupervised learning

Training examples are unlabeled, so unsupervised algorithms look for naturally occuring patterns in the data. Historically this was usually clustering which can be used for segmentation tasks like customer segmentation and anomaly detection for financial crime detection. More recently, unsupervised approaches using DNNs (Autoencoders and Generative Adversarial Networks) are being increasingly used for filtering, dimensionality reduction, and in generative applications where models generate artificial examples of training instances such as human faces.

Reinforcement learning (RL)

RL algorithms are software agents that learn policies about how to interact with their environment. They behave in a way most people expect of AI in the sense they that choose actions and 'do things' in response to other agents (such as humans). To do this, they rely on a state space representation of their environment. They seek to optimise cumulative reward over time by iteratively choosing actions that result in 'high value' states. The value of various environment states are learnt in the training phase where the algorithm explores its environment. Applications in this domain include inventory management and dynamic pricing. Google DeepMind was able to train an algorithm, AlphaGo, using RL to beat a champion Go player. This technique has also been used to train robots to climb stairs like humans and to improve lane merging software for self-driving cars.

Source: PwC

Subjective scale of explainability of different classes of algorithms and learning techniques

As discussed, explainability refers to the understandability of a given result viewed as post hoc interpretations. Since most of these models do not give direct explanations as to why or how the results are achieved, we have provided subjective values on the scale of 1 to 5 (with 1 being the most difficult and 5 being the easiest) to rate how easy or difficult it is for an end user to decipher why a model made a certain decision. Each of these learning techniques has different structures that are affected by how they learn from new information.

Exhibit 2 | Subjective scale of explainability of different classes of algorithms and learning techniques

AI algorithm class	Learning technique	Scale of explainability (1-5)	Reasoning/Explanation
Graphical models	Bayesian belief networks (BNNs)	3.5	BNNs are a statistical model used to describe the conditional dependencies between different random variables. BNNs have a high level of explainability because the probabilities associated with the parent nodes influence the final, making it possible to see how much of a certain feature is used to determine the final outcome.
Supervised or unsupervised learning	Decision trees	4	Decision trees partition data based on the highest information gain, where the nodes have the most influence. They are represented as a treelike structure where the most important feature is at the top, with other features branching off beneath it in order of relative importance. Out of all the ML learning techniques, decision trees are the most explainable because you can follow the progression of branches to determine the exact factors used in making the final prediction.
Supervised or unsupervised learning	Logistic regression	3	The most commonly used supervised learning technique, logistic regression represents how binary or multinomial response variables are related to a set of explanatory variables (that is, features). Because an equation is associated with the predictions of this model, we can investigate the influence of each feature on the final prediction. Equations can get messy, so we believe this technique is only moderately explainable.

AI algorithm class	Learning technique	Scale of explainability (1-5)	Reasoning/Explanation
Supervised or unsupervised learning	Support vector machines (SVMs)	2	SVMs are based on the concept of decision planes that define decision boundaries. SVMs are similar ``a partition, it is difficult to decipher what features were important in calculating that result.
Supervised or unsupervised learning	K-means clustering (unsupervised)	3	K-means clustering is an unsupervised learning technique used to group data into 'K' clusters of similar features. This technique is moderately explainable because one can view the centre of the clusters as descriptors of what each group represents—although, it is not always clear what certain clusters mean purely based on the centroids (centre point) of the clusters.
Deep learning	Neural networks (NNs)	1	Neural networks are the building blocks of all deep learning techniques and are becoming more prevalently used in solving ML tasks. NNs are based on the biological structure of the brain in which neurons connect to other neurons through their axon. In the same way, these networks have hidden layers of nodes where information is transferred based on the node's activation. This algorithm is the least explainable because each hidden node represents a non-linear combination of all the previous nodes. However, Israeli computer science professor Naftali Tishby's recent theoretical work in this area may help explain why and how they work.
Ensemble models	Random forest/ boosting	3	Random forest techniques operate by constructing a multitude of decision trees during training, then outputting the prediction that is the average prediction across all the trees. Even though decision trees are pretty explainable, random forest adds another layer of tree aggregation that makes understanding the final result more difficult.
Reinforcement learning (RL)	Q-learning	2	Q-learning is a technique that learns from positive and negative rewards. It uses these rewards to estimate future returns based on taking a certain action in a certain state. This technique is not very explainable because the only information given with the certain predicted action is the estimated future reward. Users would not be able to understand the agent's intent because it is looking multiple steps ahead.
Natural language processing (NLP)	Hidden Markov models (HMMs)	3	HMMs can be represented as the simplest dynamic Bayesian network (BBNs that change over time). It is a stochastic model used to model randomly changing systems where the future state only depends on the current state. Similar to BBN, you attribute certain weight to the previous states based on the probabilities, although the stochastic nature of HMM makes it slightly less explainable.

Source: PwC

Contact us for more information

Authors

Chris Oxborough
Partner
Emerging and Disruptive Technology Risk

E: chris.oxborough@pwc.com
T: +44 (0)7711 473199

Euan Cameron
Partner
UK AI Leader

E: euan.cameron@pwc.com
T: +44 (0)7802 438423

Andy Townsend
Senior Associate
Emerging and Disruptive Technology Risk

E: andrew.townsend@pwc.com
T: +44 (0)7446 893328

Dr. Anand Rao
Partner
Global AI Lead

E: anand.s.rao@pwc.com
T: +1 (617) 633 8354

Christian Westermann
Partner
Data & Analytics Leader

E: christian.westermann@ch.pwc.com
T: +41 (58)792 4400

Middle East Subject Matter Experts

Matthew White
Partner
Digital Trust Leader

E: matthew.white@pwc.com
T: +971 (0)56 113 4205

Oliver Sykes
Partner
Digital Trust

E: oliver.sykes@pwc.com
T: +971 (0)56 480 2447

Clement Chan
Director
Digital Trust

E: clement.chan@pwc.com
T: +971 (0)50 152 3619

190131-142122-PB-OS