Bringing Transparency to Machine Learning Models & Predictions

The use of artificial intelligence (AI) opens up opportunities for most industries in terms of promising breakthroughs in speed, scalability, decision-making and product personalization. The application of AI has repeatedly driven fundamental changes in processes or otherwise along the value chain of companies across the world: product development, profiling & personalization, fully automated data processing. As speed, quality and cost factors improve, customer expectations re-adjust, motivating a perpetual cycle of innovation. Adoption of AI can also create conflict between innovation, strategy, regulation as well as compliance violations.

Case in point: the General Data Protection Regulation (GDPR) entitles consumers to demand explanations of algorithmic decisions applied to them. Where neural networks power the decision engine, a multitude of parameters, layers and nodes can make this a non-trivial task. The recently proposed Artificial Intelligence Act (AI Act) raises the pressure: AI systems that directly or indirectly affect natural persons must not do so without their knowledge. Developers of AI will be required to make declarations of conformity to these standards, and, in cases of high-risk applications, base these on independent assessments conducted by independent third parties.

With this article, we illustrate the possibilities and relative effectiveness in applying so-called "explainable AI" techniques to render algorithmic decisions more comprehensible and thereby fulfil growing regulatory requirements.

Regulatory Challenges in The Use Of Artificial Intelligence (AI).

Based both on experience from GDPR, the principal regulation in Europe and UK since May 2018, as well as indications from the newly proposed AI Act (April 2021), we can identify some challenges:

- Article 6 of the AI Act prescribes stringent quality and conformity requirements on those AI systems deemed high-risk, targeting specific applications in Annex III (e.g. autonomous vehicles, credit scoring, medical imaging):
  - Risk Management system: a risk management process repeatedly updated throughout the entire lifecycle of the AI system. Residual risks limited to those judged acceptable even under conditions of possible misuse, to be communicated to users.
  - Data governance: training, validation and test on high-quality data, subject to appropriate governance. This includes a prior assessment of the availability, quantity and suitability of the data sets, as well as examination of possible biases.
  - Documentation and record-keeping: technical documentation prior to launch that demonstrates implementation to requirements such that National Competent Authorities (NCAs) may verify compliance.
  - Robustness, accuracy and security: accuracy, robustness and cybersecurity fitting the system envisioned purpose. Where needed, measures to prevent and deal with attempts to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws must be put in place.
- Article 52 of the proposed AI Act outlines transparency requirements – not exclusively for high-risk AI systems – both in terms of their interpretability as well as awareness
  - Transparency: systems designed to enable users to interpret the outputs and use them appropriately. Documentation and instructions must be complete, pertinent and easily available to users.
- Article 22 GDPR defines that - apart from exceptions - decisions regarding individuals may not be made entirely by algorithms.
- Article 35 (1) of the GDPR requires the performance of an impact assessment for processing operations for the protection of personal data when AI is used. However, a "black box" AI does not allow for such an assessment, after which the requirement does not seem to be fulfilled.
- Art. 13(2)(f) GDPR and Art. 14(2)(f) GDPR require ensuring sufficient transparency about the use of personal data, the resulting obligation to provide information leads to increased effort for companies.
- Further challenges arise from Art. 15 to 21 GDPR (rights of access, rectification, deletion). In addition, general equal treatment laws such as discrimination prohibitions must be taken into account.

Outside of Europe (e.g. in Japan), the Personal Information Protection Commission has assumed supervisory, regulatory and arbitration functions in the data protection environment since 2016. In Singapore, the Personal Data Protection Commission was installed as an authority to enforce the applicable data protection law. In California, the California Consumer Privacy Act (CCPA) came into force on January 1, 2020.

The financial service industry is accustomed to regulatory oversight. In modernizing their competitive offering with AI, institutions must take existing hurdles such as GDPR as well as national laws (on discrimination, equal treatment) into account. They must also be prepared for new directives and laws taking shape, such as the ePrivacy Directive, or safe-harbour agreements on data exchange between the USA and Europe. Already a burden on cost-efficiency, regulatory burden looks only to increase.

Explainable Algorithms Are Critical To Making AI Trustworthy

If a solution consistently and reliably works, do we necessarily need to understand how, or even why? A litany of regulations call for explainable AI, as if this were more important than the myriad benefits – dramatic enhancement of capabilities, comfort of personalization, and economies from automation. Yet there are compelling reasons to understand the inner workings of AI models – beyond checking the compliance box. Understanding is critical to the developer – without it, our only chance for improving AI is by near-random tinkering. Understanding is critical to risk management – otherwise, we may not be able to anticipate failure modes, or know the boundaries within which application of the algorithm is appropriate. As AI practitioners and responsible members of society, we need a set of techniques for use during or following the modelling stage that allow us to interpret outcomes, independently validate them, and then – and only then – place our trust in them and their increasingly dominant role in economy, science and society. To trust, we must be able to derive the algorithm's rationale, to characterize its strengths and weaknesses, and to delineate the boundaries within which they will operate in the future. We must be able to distinguish a function performing for the right reasons, on the right inputs, and not just by circumstance.

Trust, regulatory compliance, the risk of reputational damage and more efficient human-machine synergy are the reason why trustworthiness is gaining momentum for machine learning deployment. However, algorithm interpretability and trustworthiness of an AI system are not synonymous; touting explainable AI as a cure-all to garner trust is grounds for caution. Many approaches exist to explain algorithms, often suitable only under a narrow set of conditions. The danger is that they – even when incorrectly applied – provide an answer, an interpretation – even if it is out of context and incorrect. Potential results range from a false sense of security to frustration where there is no cause.

> Trust, regulatory compliance, the risk of reputational damage and more efficient human-machine synergy are the reason why trustworthiness is gaining momentum for machine learning deployment.

Moreover, how we approach explainability is as much a question of policy as it is of technology (analysis techniques). What constitutes a sufficiently interpreted AI? Is it enough to identify the decision drivers? Must we also know why they contribute? Must we be able to judge relative model strengths and limitations? Different explanation needs apply depending on the use case, the regulation in question and even on the user of the AI model or system. Context is key when seeking insights about an AI model. While the developer might focus on the stability and accuracy of the models, the operator or the subject matter expert would want to understand how an output is derived and whether his or her approach aligns with the machine. On the other hand, the management overlooking the models would rather target the costs required to maintain the models whereas the regulator might seek insight about the overall risks related to the techniques.

Using the right tools at the right situation and combining the provided insights efficiently to satisfy all the interests from different context are challenges posed by the field of explainable AI.

In order to address such challenges, explanations will vary depending on their target audience. Some will be simple, others highly technical, detailed and requiring a strong mathematical background. To illustrate this range of possibilities, we present four state-of-the-art explanation approaches [1]:

- User benefit: designed to inform a user about an outcome. For example, the reason why a loan application was approved or denied to the applicant.
- Societal acceptance: designed to generate trust and acceptance by society. For example, if an AI system produces unexpected results, the explanation may help users understand how and why the system came to the counter-intuitive conclusion. Providing a rationale may also provide an increased sense of comfort in the system.
- Regulatory compliance: assists audits of compliance or adherence to safety standards. The target audience may include users seeking details (e.g., a safety regulator) or those interacting with the system (e.g., a developer).
- System development: assists or facilitates developing, improving, debugging, and maintaining of an AI algorithm or system. The target audience includes: technical staff, product managers, and executives. Such users require significant detail to determine root cause, not merely association or correlation.

AI Algorithms

We can start talking about explainable algorithms by introducing “self-explanatory models”, that include algorithms capable of inherently illustrate the derivation of their conclusions. The user gains sufficient understanding from viewing outputs and querying the models.

The generalisation of such algorithms to interpret other AI models, has been the first type of model explanation used by developer. However, as of now, we do not only rely on the generalisation of self-explanatory model but also on techniques built ad-hoc for explanations purpose. In general terms, model interpretation takes two forms:

- Global explanations: fundamental workings of a model across all decision outputs
- Local explanations: distinct derivation of each individual model output, each decision. Counterfactual explanations are a subset of local explanations, describing the “what if” scenario, or what the model would have predicted on different inputs. Counterfactuals essentially measure the distance to the decision boundary.

Self-explanatory models include simple algorithms such as Decision Trees, Linear and Logistic Regression. While they do inherently provide insight, prediction results are generally less accurate than complex models such as neural networks. For example, one possible interpretation of self-explanatory models like linear or logistic regression, is using the weights of their coefficients to indicate the importance of features. However, they can be inaccurate when the data is non-linear. One indicator of the relative interpretability of a regression model is the number of non-zero coefficients, as sparser model are generally considered more understandable.

The limitations of such simple algorithms have encouraged research into more accurate self-explanatory models. In the explainable AI community, it is often argued that using models that explain themselves is the best way to produce comprehensible models, as separately produced explanations of black-box models may not be faithful to what the original model computes. The reason behind such a claim is that explanations often have low explanation accuracy if those explanations are not the models themselves.

A recently developed, more general class of self-explanatory model compensates for the high bias of the regression model: Generalized Additive Models (GA2M). These models allow for the introduction of non-linear behaviour in the model parameters, dropping the high bias assumption of linear and logistic regression while keeping the model explainable and sparse.

Global Explanations

AI models that do not provide a (meaningful) explanation can be interpreted using alternate algorithms. This approach essentially builds an interpretable surrogate model that mimics the original model – perhaps less accurate, but directionally correct and substantially more interpretable. Therefore, the quality of gained information through these surrogate models depend on many different factors (e.g. choice of models, quality of input data, interpretation methods).

With global explanation, we seek a general understanding of the algorithm's internal logic across all predictions. These follow an approach of querying the black-box algorithm from different angles in order to study its response and derive a comparable, simpler surrogate model that lends itself well to explanation.

> With global explanation, we seek a general understanding of the algorithm's internal logic across all predictions.

The most renowned global explanations algorithm is SHAP (SHapley Additive exPlanations) built on principles of game theory. SHAP provides a global per-feature importance for a regression problem by converting it to a coalitional game [1]. Coalitional games portray a dynamic where n players may form varying coalitions in order to improve collective chances of winning and divvy up the resulting payoff. (Statistically the total payoff is often largest when all players team up.) One mathematically appealing ("fair") scheme to allocate payoffs in is to award players according to their relative contribution to the game, mathematically represented by their Shapley value. SHAP applies this methodology to the regression outputs of an AI model, where the model features are the players and the prediction target is the payoff [1]. Features (players) either participate or do not participate in coalition with the other features for each data record processed. SHAP then computes the Shapley values for each feature and considers them as representations of feature importance.

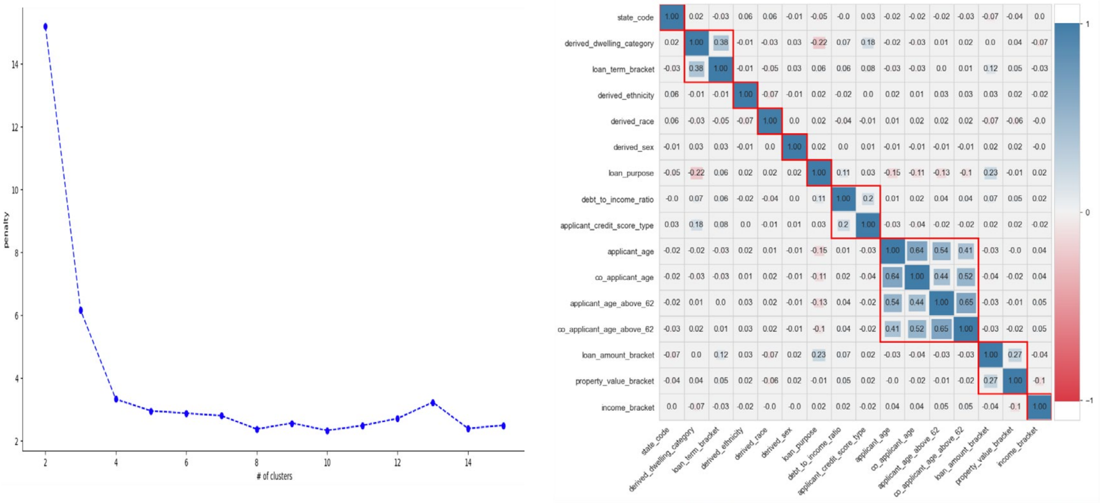
The currently available open-source implementation of SHAP relies on the assumption that the features under consideration are entirely independent without interaction. Where this may be reasonable for the coalition game analogy of players, it is often not applicable in real-world study of machine learning models. Two unfortunate effects arise from this simplification assumption:

1. feature interaction is categorically overlooked, leaving one key question unanswered

2. the ranking of feature importance is skewed; the reported explainability output may mislead

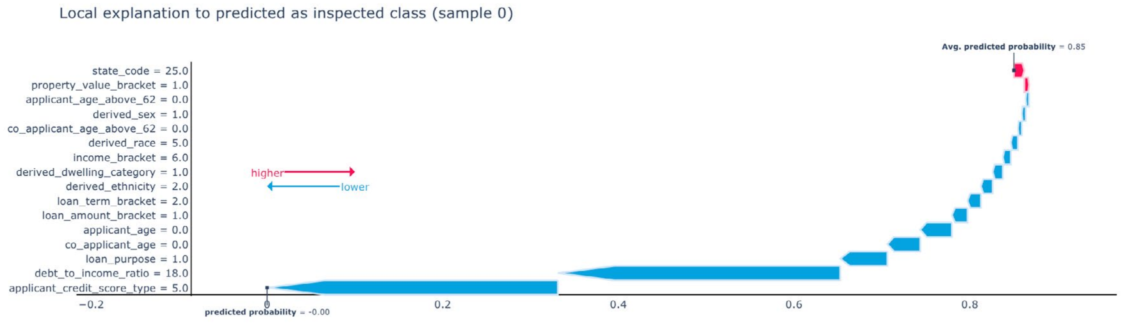
It is possible to overcome these limitations, to explain models without introducing possible bias and inaccuracy into the explanations themselves. Novel research from Redelmeier et al. [2], explores how SHAP values could be calculated without reliance on the independency assumption. This approach makes use of Conditional Trees instead of a regression model. The technique is more computationally intensive, yet provides a significantly deeper inspection into the model's inner-workings, notably how multiple features interact and a correct rendition of the respective importance taking those interactions into consideration. Deloitte's AI explainability tool Lucid[ML] implements this cutting-edge technique (there is no open-source equivalent available at the time of writing). This more advanced implementation produces outputs as depicted in Fig. 1:

Fig. 1 – Feature clustering and local explanation with clustered feature.
The novel implementation of SHAP values allows for the study and contribution by clustered feature, removing possible bias and assumptions on the model, providing more accurate explanations.



	state_code	derived_dwelling_category	loan_term_bracket	derived_ethnicity	derived_race	derived_sex	loan_purpose	debt_to_income_ratio	applicant_credit_score_type	applicant_age	co_applicant_age	applicant_age_above_62	co_applicant_age_above_62	loan_amount_bracket	property_value_bracket	income_bracket
state_code	1.00	0.02	-0.03	0.06	0.06	-0.01	-0.05	-0.0	0.03	-0.02	-0.02	-0.02	-0.03	-0.07	-0.04	0.0
derived_dwelling_category	0.02	1.00	0.38	-0.01	-0.03	0.03	-0.22	0.07	0.18	-0.02	-0.03	0.01	0.02	0.0	0.04	-0.07
loan_term_bracket	-0.03	0.38	1.00	-0.01	-0.05	0.03	0.06	0.06	0.08	-0.03	-0.03	0.0	0.01	0.12	0.05	-0.03
derived_ethnicity	0.06	-0.01	-0.01	1.00	-0.07	-0.01	0.02	-0.02	0.0	0.02	0.01	0.03	0.03	0.03	0.02	-0.02
derived_race	0.06	-0.03	-0.05	-0.07	1.00	0.0	0.02	-0.04	-0.01	0.01	0.02	0.02	0.02	-0.07	-0.06	-0.0
derived_sex	-0.01	0.03	0.03	-0.01	0.0	1.00	0.02	0.0	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.02	-0.0
loan_purpose	-0.05	-0.22	0.06	0.02	0.02	0.02	1.00	0.11	0.03	-0.15	-0.11	-0.13	-0.1	0.23	-0.01	0.02
debt_to_income_ratio	-0.0	0.07	0.06	-0.02	-0.04	0.0	0.11	1.00	0.2	0.01	0.02	0.04	0.04	0.07	0.05	0.02
applicant_credit_score_type	0.03	0.18	0.08	0.0	-0.01	0.01	0.03	0.2	1.00	-0.03	-0.04	-0.02	-0.02	0.02	0.02	-0.02
applicant_age	-0.02	-0.02	-0.03	0.02	0.01	-0.01	-0.15	0.01	-0.03	1.00	0.64	0.54	0.41	-0.03	-0.0	0.04
co_applicant_age	-0.02	-0.03	-0.03	0.01	0.02	-0.01	-0.11	0.02	-0.04	0.64	1.00	0.44	0.52	-0.04	-0.02	0.04
applicant_age_above_62	-0.02	0.01	0.0	0.03	0.02	-0.01	-0.13	0.04	-0.02	0.54	0.44	1.00	0.65	-0.03	-0.01	0.05
co_applicant_age_above_62	-0.03	0.02	0.01	0.03	0.02	-0.01	-0.1	0.04	-0.02	0.41	0.52	0.65	1.00	-0.03	-0.02	0.05
loan_amount_bracket	-0.07	0.0	0.12	0.03	-0.07	0.02	0.23	0.07	0.02	-0.03	-0.04	-0.03	-0.03	1.00	0.27	-0.04
property_value_bracket	-0.04	0.04	0.05	0.02	-0.06	0.02	-0.01	0.05	0.02	-0.0	-0.02	-0.01	-0.02	0.27	1.00	-0.1
income_bracket	0.0	-0.07	-0.03	-0.02	-0.0	-0.0	0.02	0.02	-0.02	0.04	0.04	0.05	0.05	-0.04	-0.1	1.00

Fig. 2 – Waterfall Plot.
Relative importance – and directional contribution – is most easily viewed via a waterfall chart, a useful supplement to the feature correlation & cluster matrix.



Partial Dependence Plots (PDPs) are another effective global explanation technique. They show the marginal change of the predicted response when a feature (specifically: a value of that specific data field or component) changes. PDPs are particularly useful to determine whether a relationship between a feature and the response is linear or more complex [1].

Local Explanation

With local explanation, we seek to understand the rationale of a model as pertains to a prediction / decision outcome for a single case, a single data record. For example, if a classifier model is used as a sophisticated, multivariate credit scoring decision tool, we may be required under GDPR to explain to consumers why their loan application had been rejected. Local differ from global explanations as they need only to explain a single outcome, requiring no generalization to other decisions.

> With local explanation, we seek to understand the rationale of a model as pertains to a prediction/ decision outcome for a single case, a single data record.

One popular local explanation technique is LIME (Local Interpretable Model-Agnostic Explainer). LIME works by querying nearby points to construct an interpretable model that represents only the portion of the overall classifier relating to that individual decision. It then uses local approximate model to provide insights on the original model logic. By default, LIME makes use of logistic regression to construct the local approximate model, which introduces limitations on the fidelity and original model complexity that LIME can approximate.

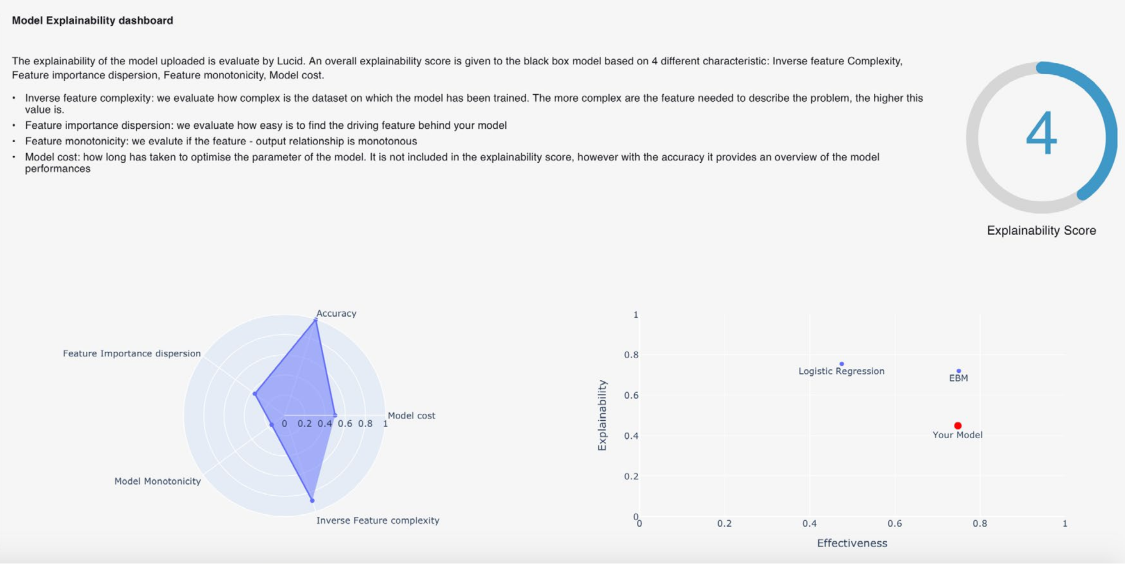
Another popular method is by way of counterfactuals, the "what if" scenarios of alternative inputs that could possibly lead to alternative outcomes. An example of an informative counterfactual analysis would be to determine the minimum amount an input must to change in order to for the system's output (prediction / decision) also to change.

Wachter et al. [3], were the first to introduce the concept of counterfactual analysis, by measuring the distance of the counterfactual variant from the original data point. One major limitation of counterfactual analysis is its dependency on the model under investigation. Their loss-functions do not support tree-based models widely in use.

Better Counterfactual Analysis

The counterfactual approach developed for Deloitte's explainable AI tool Lucid [ML] overcomes these limitations, providing a truly model-agnostic local interpretation approach. Furthermore, Lucid's approach enables the user to progressively refine the counterfactual analysis by specifying which features may be changed, which may not, or which may be changed only under certain conditions. This added flexibility permits embedding SME knowledge into the counterfactual generating algorithm itself, paving the way for deeper analysis.

Fig. 3 – Deloitte Lucid tool: explainability dashboard.
The explainability of a model is evaluated with a score between 0 and 10, according to 3 dimensions: feature sparsity, feature monotonicity and feature complexity. The model under study is also compared to less complex model (logistic regression in this case) and to a surrogate white box model.



Model Explainability dashboard

The explainability of the model uploaded is evaluate by Lucid. An overall explainability score is given to the black box model based on 4 different characteristic: Inverse feature Complexity, Feature importance dispersion, Feature monotonicity, Model cost.

- Inverse feature complexity: we evaluate how complex is the dataset on which the model has been trained. The more complex are the feature needed to describe the problem, the higher this value is.
- Feature importance dispersion: we evaluate how easy is to find the driving feature behind your model
- Feature monotonicity: we evalute if the feature - output relationship is monotonous
- Model cost: how long has taken to optimise the parameter of the model. It is not included in the explainability score, however with the accuracy it provides an overview of the model performances

The Explainability Trade-Off

We have seen that improving transparency of AI models is a multi-faceted exercise requiring some care and sophistication to yield useful results. General model interpretability and derivation of individual decision rationale are not the same question. Motivations to understand at global and local levels are likewise distinct – from de-bugging to defending a model. The common theme is to ensure model predictions do not lead us astray and model-supported decisions may be properly accounted for.

When designing an AI model, we iteratively tune it to maximize performance, often using prediction accuracy as our sole gauge of quality, or how well it identifies patterns that we humans would overlook. Where hyperparameter tuning may increase predictive power, this may come at a cost of complexity and reduce our ability to interpret its internal function. The trade-off between predictive power and model transparency is made even more difficult in the lack of straightforward metrics for explainability itself. We propose resolving this through introduction of a multi-dimensional metric to avoid introducing an unfavourable design bias:

1. Feature sparsity: the fewer features a model requires, the easier it is to understand.

2. Feature monotonicity: as clear and direct as possible a relationship between input and output.

3. Feature complexity: the minimum features/data required to describe the problem.

Based on a weighted score of these three dimensions, we present an explainability score which we compare to the accuracy of the model, in order to better understand that trade-off (fig. 2).

Conclusion

XAI ("Explainable AI") is an active area of research with a colourful array of methods seeking to cast light into black box machine learning models. The unique motivations and challenges surrounding model explainability at a global or at a local level each require dedicated approaches to provide any satisfactory result. These approaches also require perspective to prevent application out of context and thereby risking misinterpretation of models. Just as a three dimensional object can only truly be perceived by viewing at different angles, models can only truly be interpreted by applying a comprehensive set of techniques, each within its boundary of applicability.

Artificial intelligence must be transparent in order to gain widespread acceptance, winning the trust of the full spectrum of stakeholders – developers, subject-matter experts, management, auditors, regulators, employees and customers. XAI methodologies and tools can play a central role toward achieving this acceptance – as well as to improving the quality of AI through the additional transparency and understanding. Highly regulated industries, such as banking, insurance, pharmaceuticals and automotive will all benefit from explainable AI, enabling innovation while managing liability risk and ensuring compliance to the regulatory terrain within which they operate. As AI becomes more transparent and comprehensible, quality and reliability will improve, use-cases will proliferate and its power to transform industry, science and economy will expand.

Contacts

David Thogmartin
Director
aiStudio | AI & Data Analytics
dthogmartin@deloitte.de

www.deloitte.com/de/aistudio

References

[1] Phillips et al., doi.org/10.6028/NIST.IR.8312-draft
[2] Redelmeier et al. [1], arXiv:2007.01027v1
[3] Wachter et al., arXiv:1711.00399

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/de/UeberUns to learn more.

Deloitte is a leading global provider of audit and assurance, consulting, financial advisory, risk advisory, tax and related services; legal advisory services in Germany are provided by Deloitte Legal. Our global network of member firms and related entities in more than 150 countries and territories (collectively, the "Deloitte organization") serves four out of five Fortune Global 500® companies. Learn how Deloitte's approximately 330,000 people make an impact that matters at www.deloitte.com/de.

This communication and any attachment to it is for internal distribution among personnel of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms and their related entities (collectively, the "Deloitte organization"). It may contain confidential information and is intended solely for the use of the individual or entity to whom it is addressed. If you are not the intended recipient, please notify us immediately and then please delete this communication and all copies of it on your system. Please do not use this communication in any way.

None of DTTL, its member firms, related entities, employees or agents shall be responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. DTTL and each of its member firms, and their related entities, are legally separate and independent entities.

Issue 06/2021