The three stages of Explainable AI: How explainability facilitates real world deployment of AI

Clodéric Mars*, Rémi Dès*, Matthieu Boussard*

*craft ai
8, rue de la Michodière, 75002 Paris
prenom@craft.ai
https://www.craft.ai

Résumé. L'intérêt pour l'IA explicable s'est récemment vu renouvelé, et nous pensons que ces approches permettent de faire une vraie différence dans le déploiement d'IA, particulièrement dans le monde de l'entreprise. Dans cet article nous introduisons un cadre permettant de catégoriser les niveaux d'explicabilité, leurs impacts dans l'opérationnalisation d'IA, et leurs prérequis.

1 Introduction

The explainability of AI has become a major concern for AI builders and users, especially in the enterprise world. As AIs have more and more impact on the daily operations of businesses, we see trust, acceptance, accountability and certifiability become requirements for any deployment at a large scale.

1.1 XAI

Explainable AI (XAI) as a field was popularized by the eponymous DARPA program launched in 2017, with the goal of creating a suite of machine learning techniques that produce more *"explainable"* models while maintaining a high level of learning performance, thus enabling human users to understand, trust and effectively manage the emerging generation of AI systems (Gunning, 2017).

Over the past 5 years, explainability has become a key part of the AI industry strategy for countries (Holdren et Smith, 2016; Villani et al., 2018) or research institutions (Braunschweig, 2016). It is also a strategic axis for companies, small (Mars, 2019; Guggiola et al., 2018) or large, for example through the publication of open source libraries (Microsoft, 2019; IBM, 2019) or dedicated services (Google). Major AI conferences dedicate workshops to this topic (e.g. First Workshop on Explainable Artificial Intelligence (XAI) IJCAI 2017). Specialized MOOC are being launched (Becker, 2019).

1.2 What is an explanation?

Explainable AI is about providing explanations regarding AI processes to stakeholders, it is therefore interesting to look at how people explain their decisions to each others. Specifically, the design of XAI can benefit from the learnings of social sciences on explanation. In his thorough review, Miller (Miller, 2019) studied works from various branches of social sciences from philosophy to cognitive science and psychology.

The surveyed body of work tends to show that people seek to build a mental model of how decisions are made or how events occur, in order to anticipate them and reason about them. Explanations are a way to build such models much quicker than through observation only. Because mental models are inherently subjective, good explanations are biased towards the explainee to match their perspective and their preexisting knowledge. In the real world examples we describe below, we found that the work of understanding the point of view of the explainee is a major part of the design of explainable AIs.

Another major finding is that good explanations are contrastive. It is not about answering *"why has event E occurred?"* but rather *"why has event E occurred instead of another event C?"*. We found out that the capability to generate such *constrative* or *counterfactual* explanations is quite important in the deployed systems we describe in section 2.2.

Miller argues that Explainable AI as a field should be considered at the crossroad of Social Science, Human-Computer Interaction and Artificial Intelligence. Taking a more practical approach, in this article we will take the point of view of the people and systems interacting with AI systems, and study how explainability impacts these interactions in terms of features, acceptance and capacity to be deployed.

2 How XAI makes a difference

In order to study the impact XAI makes on AI projects we are categorizing effects in three stages described in Fig 1. Higher stages require higher levels of explainability and have more impact on the resulting AIs. We take the point of view of the industrial world, and look at how explainability can make a difference in the deployment and application of AI.

This work is based on the experience we gathered working and discussing with our customers, partners and community, as a provider of machine learning solutions. Examples are focused on systems based on Machine Learning but the proposed three stages are relevant to any kind of AI.

2.1 Stage 1: Explainable building process

In any organisation, just like any IT project, a project leveraging AI aims to have an impact on the daily job of some people. Its goal might even directly be to automate part of worker's job or to help them deliver value they could not before. Especially when AI is involved, affected users can be wary of the new system. In particular they may feel threatened by the automation of some of their tasks, or may not believe that a simple *computer program* can execute complex tasks correctly. A recommandable method to address those concerns is to involve them in the building of the AI. This is where explainability plays a big role.

In this context, traditional quality metrics such as confusion matrices, r2, RMSE, MAE, etc. are not sufficient to get the future AI user's trust, since they want to know more about the *why* than about the raw results. Visualization is the first go-to technique. Simply plotting the output against context variables is a good way to get a *feel* for how an AI performs over the target domain when dimensionality is low. Interactive simulations can help explore the

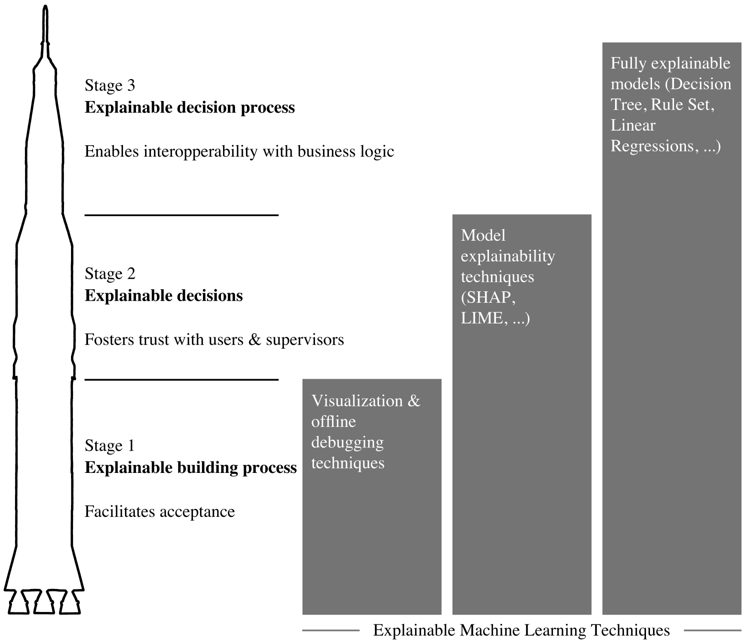
Stage 3
Explainable decision process
Enables interopperability with business logic
Stage 2
Explainable decisions
Fosters trust with users & supervisors
Stage 1
Explainable building process
Facilitates acceptance
Visualization & offline debugging techniques
Model explainability techniques (SHAP, LIME, ...)
Fully explainable models (Decision Tree, Rule Set, Linear Regressions, ...)
Explainable Machine Learning Techniques

FIG. 1 – *The three stages of AI explainability*

domain to experience how the AI will react. Beyond these techniques which are applicable to any *black box* computations, more advanced techniques open the hood and make the structure of the AI itself inspectable.

In the following sections describing the subsequent stages, we will talk about techniques able to work while the AIs are *live*, processing production data, at production speed. These techniques are also well suited for stage 1, where the inspection is offline, with less data and runtime constraints.

Debugging tools that were initially designed for data scientists can also be leveraged for other stakeholders. AIs powered by neural networks can be inspected by visualizing how intermediate layers *react* to different input, Tensorflow Playground (Smilkov et Carter) or ConvnetJS (Karpathy, 2014) are good examples of this approach. On images, the computation of *saliency maps* can also help to convey which parts of the image are considered by the network to make its prediction (Simonyan et al., 2013). This technique led to the identification of the infamous *husky vs wolf* issue in which a wolf is primarily identified by the presence of snow in the picture (Ribeiro et al., 2016b). Tools like Seq2Seq Vis (Strobelt et al., 2018) bring the same kind of debugging capabilities to natural language focused neural networks. This shows that even neural networks, which are considered black boxes, can be at least partly explained offline to the non-technical AI project stakeholder by using the right tools.

While the initial goal of explaining why the AI works the way it does is to ease its adoption, explainability also increases the involvement of potential users by letting them achieve a deeper understanding. As a result they can assist in its development, ensuring that the AI solves an actual problem, and provide valuable feedback on specific behaviors of the AI: instead of providing knowledge upfront, it is always easier to *react* to what you see the AI doing and why it does it. In many cases, domain experts can easily help if they have an understanding of why the AI makes decisions: sensors having an undocumented validity domain, well-known contexts leading to corrupted data, spurious correlations because of a missing data sources, etc.

The first stage of explainability is about helping create a multi disciplinary team of experts in their respective fields who understand the AI they are building. Offline explainability techniques are key to the acceptance of the future AI and create opportunities to build a better system.

2.2 Stage 2: Explainable decisions

Trust in a system is key, especially in an enterprise tool that has an impact on day to day business. Trust makes the difference between a system that is *"micro managed"* by its users or supervisors, and a system that can enjoy a larger autonomy. The more management a system needs, the more manpower it requires and therefore the less value it has.

Trust is built when a system is not surprising, when it behaves according to our mental model. A system whose limits are understood by its users is arguably more valuable than a more accurate system whose results are considered unreliable. As discussed in section 1.2, explanations are a good way to accelerate the construction of this mental model. That is where the capacity to explain the AIs' decisions has an impact. That is the second stage of explainable AI.

Stage 1 explainability does not have the same impact: most users or supervisors of AIs did not have the chance to participate in their inception, and in more and more cases, AIs can evolve over time. Furthermore, the ability to access explanations of past AI decisions can help pinpoint root causes and generally provide traceability.

The ability to provide explanations to any AI decision is an active field of innovation with methods such as TreeInterpreter (Saabas, 2014), LIME (Ribeiro et al., 2016a) or SHAP (Lundberg et Lee, 2017). Given a predictive model and a prediction, these methods aim at providing a local explanation for the prediction. This explanation takes the form of linear factors that can be applied to input features to reach the predicted results, thus giving an idea of the local feature importance and behavior of the model. The computed feature factors can also be used to generated counterfactual examples and give an idea of the trend of the predicted value given changes in the input features.

An interesting property of this class of algorithm is that they can work using a feature set that is different from the actual feature set used by the model. It is therefore possible to adapt the explanation by making it more comprehensible to the explainee, independently from the features that yield the best predictions. This additional feature engineering step is not without risk, as it can be used to convince explainees to blindly trust said AI, by presenting a deceptive approximation instead of bringing more transparency (Denis et Varenne, 2019).

A good example of SHAP usage can be found in the banking fraud detection solution (Mars, 2019) provided by the Bleckwen company. One key part of the solution is a predictive model, trained on labelled datasets containing fraudulent and non-fraudulent transactions. This model computes a score for each transactions. Transactions having a score above a certain threshold are reviewed by a human expert to confirm their fraudulent nature. One of their customers' requirements is to get explanations for every score. They chose to use non-explainable gradient boosting techniques for the model on a range of complex features. The local explanation is computed by SHAP on a range of features they designed with their end users to make them completely understandable to them.

Another example of stage 2 explainable AI is how Dalkia uses machine learning as a part of their energy management dashboard. Here, decision trees are used to predict an energy diagnosis based on labelled data streams. Predictions are used as diagnosis recommendations in the energy managers' dashboard, and explanations are extracted from the decision tree as a set of rules that were applied (craft ai, 2018). What's really interesting in this example is that without explanations alongside the recommendations this AI would not have any value. At its core, the goal of the system is to help energy managers handle more data points. Without an explanation, when provided with a prediction, energy managers would need to investigate the raw data in order to confirm or contradict it. They would end up doing the same amount of work as without explanations. When an explanation is provided, this counter investigation is only needed when the energy manager disagrees with it. Here, explanations are needed for the business value of the AI.

2.3 Stage 3: Explainable decision process

Stages 1 and 2 are about helping humans create a mental model of how AIs operate. This enables humans to *"reason"* about the way AIs work critically, and decide when to trust them and accept their outputs, predictions or recommendations. To scale this up to many AIs and over time, you need to define business logic that will apply the same *"reasoning"* automatically. Stage 3 is about enabling interoperability between AIs and other pieces of software, especially software that uses business logic.

When discussing AI, and especially models generated through machine learning, we often talk about the underlying concepts they capture, for example convolutional neural networks are able to recognize visual patterns and build upon these lower level *"concepts"* in their predictions. AIs that can explain those lower level building blocks, make them inspectable to business logic, reach stage 3. Such AIs ultimately act as a knowledge base of the behavior they model.

Stage 3 explainability makes a difference especially when a lot of instances of evolving AIs need to be supervised by business logic, for example in a context of continuous certifiability or collaborative automation between machine learning based AIs and business rules.

This level of explainability requires fully explainable AI. Machine learning techniques such as linear regressions or decision tree learning (Quinlan, 1993) can reach such levels. Another approach is to approximate a more *"black box"* model with a more explainable model, for example RuleFit is able to learn a minimal ensemble of rules from a tree ensemble method such as Random Forest (Friedman et al., 2008).

An interesting example of level 3 explainability is Total Direct Energie's energy coaching feature that is part of their customer-facing mobile application. It generates personalized messages for each customer (craft ai, 2019). At its core, the system is made of a machine learning-based energy consumption predictive model, and a business expertise-based message generation and selection module. The predictive model is made of individual regression trees, each updated continuously from the data of a single household. The message generation module is generic for all users, and uses the model's explanations and predictions as input data to select and personalize each message. So the predictive models provide an understanding of the household's energy consumption behavior, which is automatically processed to generate personalized messages.

When presented with a visual explanation of a decision process, people tend to navigate through its structure to understand the process. Stage 3 is about letting software programs, other AIs, do the same thing, thus unlocking a wealth of additional use cases.

3 Challenges

While there are already deployed AIs covering these three stages, there are still challenges ahead before explainable AI can be generalized.

3.1 Evaluating explanations

In the previous sections we discussed how certain techniques bring more or less explainability, however we did not discuss how we can make such an assessment.

Ad-hoc experiments or KPI can be used. For example the D-Edge company, which provides pricing recommendations to hotel managers among other services, measures whether explained recommendations are accepted. Every recommendation is accompanied by a natural language explanation. Managers can accept and apply the recommendation to their pricing or discard it. As presented during a round table focused on XAI (Mars, 2019), they consider the proportion of accepted recommendations as a proxy measure for the quality of their explanation. We believe that this makes sense, as hotel managers need to be convinced to make such an impactful change to their business.

In the general case, other proxy measures can be used, such as the number of rules, nodes or input variables considered in an explanation or explainable model. However these lack generality: how can the explainability of a linear regression and of a regression tree be compared? They also lack an experimental, measurable ground truth: for example we do not know if humans find that the explainability provided by LIME grows exponentially or linearly with the number of features involved. Furthermore, as discussed in section 1.2, what constitutes a good or a bad explanation depends on the recipient of the explanation and their own cognitive biases (Denis et Varenne, 2019). This poses an additional challenge to this evaluation. There is a lack of a systemic framework or objective criteria to evaluate the explanations provided by AIs (Weller, 2017).

3.2 Improving the performances of XAI

The AI community generally considers that the more explainability you gain, the less predictive performance you can achieve, especially in Machine Learning. Overcoming this is a primary goal of the XAI field, and in particular it is the main goal of the DARPA XAI program (Gunning, 2017). Several opportunities have been identified to achieve this objective, the most promising ones being to create hybrid AI combining different approaches. One idea is to *push* high-performance but unexplainable algorithms to the edges, around an explainable core. For example in cat image recognition, a deep neural networks would identify low level details like whiskers and pointy ears, while decision trees or bayesian models would associate the presence of both whiskers and pointy ears to a cat in an explainable fashion. Another idea is to adapt

Machine Learning algorithms to work from existing expert-built symbolic representations of physical models to leverage existing knowledge, instead of having to relearn and embed it. This field is relatively new, and comes as a stark departure from the deep learning trend of the past few years.

4 Conclusion

In this paper we structured in three stages the impact that explainability can have on AI applications deployed in the *"real world"*. Those 3 stages provide a simple framework to quickly identify the need for explainability in a AI powered project. Stage 1 is about leveraging explainability to improve the adoption and performance of AI applications. Stage 2 is about explaining every AI decisions to build trust with their users and supervisors. Stage 3 is about enabling the interoperability of AI systems with each other and other software, thus unlocking new and richer use cases.

Because we focused on what explainability enables in AI, we did not discuss regulation. However it is important to note that initiatives such as the European GDPR pave the way for a *"right to explanation"* which will require, at least in some cases, a stage 2 requirement (Burt, 2017). We strongly believe that stage 2 explainability is a key to actually operationalize enterprise AI because it not only offers stronger guarantees in terms of data governance, but also facilitates involvement and support from users and domain experts impacted by such AI.

Far from being just a constraint on AI design, explainability helps develop better and richer AIs.

References

Becker, D. (2019). Machine learning explainability.

Braunschweig, B. (2016). Intelligence artificielle, les défis actuels et l'actions d'inria. Technical report, INRIA.

Burt, A. (2017). Is there a 'right to explanation' for machine learning in the gdpr?

craft ai (2018). How the deployment of an explainable ai solution improves energy performance management at dalkia.

craft ai (2019). How total direct energie applies explainable ai to its virtual assistant.

Denis, C. et F. Varenne (2019). Interpretability and explicability for machine learning: between descriptive models, predictive models and causal models. A necessary epistemological clarification. In *National (French) Conference on Artificial Intelligence (CNIA) - Artificial Intelligence Platform (PFIA)*, Toulouse, France, pp. 60–68.

Friedman, J. H., B. E. Popescu, et al. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics 2*(3), 916–954.

Google. Google explainable ai.

Guggiola, A., J.-M. Schertzer, A. Hoff, C. Ledoux, S. Monnier, O. Wautier, et H. Stalla-Bourdillion (2018). Ia explique toi ! Technical report, Quantmetry.

Gunning, D. (2017). Xai, program update november 2017. Technical report, DARPA.

Holdren, J. P. et M. Smith (2016). Preparing for the future of artificial intelligence. Technical report, Executive Office of the President National Science and Technology Council Committee on Technology. Washington, DC.

IBM (2019). Ai explainability 360. IBM.

Karpathy, A. (2014). Convnetjs: Deep learning in your browser.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774.

Mars, C. (2019). Explainable ai, a game changer for ai in production - ai night 2019 workshop.

Microsoft (2019). Interpretml.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016b). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM.

Saabas, A. (2014). Interpreting random forests.

Simonyan, K., A. Vedaldi, et A. Zisserman (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smilkov, D. et S. Carter. Tensorflow playground. Google.

Strobelt, H., S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, et A. M. Rush (2018). Debugging sequence-to-sequence models with seq2seq-vis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 368–370.

Villani, C., Y. Bonnet, C. Berthet, F. Levin, M. Schoenauer, A. C. Cornut, et B. Rondepierre (2018). Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne. Technical report, Conseil national du numérique.

Weller, A. (2017). Challenges for transparency. *arXiv preprint arXiv:1708.01870*.

Summary

Explainable AI has recently seen a renewed interest. We believe these techniques make a true difference when it comes to deploying AIs, especially in the entreprise world. In this article we introduce a framework categorizing explainability levels, their impact on operationalized AI and their requirements.