Santander X Innovation Xperts

Interpretability by design: opportunities ahead

September 2023

By Santander

Authors

Dr. Irene Unceta

Irene Unceta obtained a Bachelor's Degree in Physics from the University of Barcelona in 2012 and graduated from the MSc in Computational Science from the University of Amsterdam in 2021. She obtained her Industrial PhD from the University of Barcelona and BBVA Data & Analytics for a work on adaptation of machine learning systems in company production environments. Equal parts a Social Scientist and a Data Scientist, her professional career has been a journey through the intersection of both worlds, with a main focus on the social impact on automated decision making systems, in terms of transparency, fairness and accountability.

Table of contents

01

Introduction

Machine learning offers the promise of more accurate, efficient, and consistent information processing in a wide range of fields[1]**.** From finance and insurance, to healthcare, advertising, or the criminal justice system, this technology could lead to better decision-making processes and potentially bring substantial benefits to users, businesses, and the society at large. To deliver on this promise, however, machine learning based systems must overcome barriers to usage and adoption.

Traditional machine learning architectures, such as decision trees and rule-based systems are well-aligned with human knowledge representation processes. They can be broken down into simple, traceable rules. They also have a bounded size and include a limited number of variables. As opposed to these, current trends in model design and delivery favour fine tuning large, relatively general architectures.

Recent advances in computer vision and natural language processing have introduced mechanisms for self-attention and self-supervision, such as those in transformers, among others. Hybrid models have also gained popularity thanks to their ability to combine different deep learning structures with probabilistic measures to model uncertainty. These architectures are able to exploit nonlinearities in the data thanks to highly recursive layers of abstraction. The mathematical artifacts required to describe the relationships among such layers are more and more complex. As a result, the resulting are increasingly hard to understand. Understanding is usually at odds with size and complexity. A tendency towards increasing the number and size of layers, the intricacy and recursiveness of non-linear transformations, or the number of learners, will naturally lead to a greater opacity.

This paper defines and evaluates the issues resulting from an **increasing opacity in upcoming machine learning trends**, taking on the issue of performance (**section 2**). It then moves towards evaluating existing solutions (**section 3**), including explainability. It introduces interpretability as a potential alternative and promotes it as a most promising path forward. It identifies the key opportunities offered by interpretable models (**section 4**). It then discusses (**section 5**) the challenges for delivering interpretability by design in practice and provides key areas and sectors where these models may offer an advantage.

1. Machine learning models fall under the category of actuarial decision-making methods, which are set up on empirically established relations coded into mathematical formulas that produce automated outputs. As such, they are meant to escape the biases, distortions and beliefs related to the clinical method, which is wholly based on the human processing of information [Dawes et al., 1989].

02

The challenge posed by complexity and opacity

Independently of their specific architecture, machine learning models can be hard for humans to comprehend. The logic that enables models to produce an output based on a data input is often not well understood by users, who generally perceive them as opaque. This *opacity* can be intentional[2]. Industrial machine learning models can be subject to corporate or state secrecy in companies' legitimate interest to gain and maintain a competitive advantage[3]. Opacity can also arise because of technical illiteracy. The population at large lacks the skills required to read through code[4] or follow the complex mathematical derivations behind most algorithms. In many cases, however, **opacity is inherent to machine learning**. To extract insights from large volumes of high-dimensional data, models often possess a degree of complexity that is at odds with the demands of human-scale reasoning[5]. This is especially true for deep learning-based models and, to a lower extent, for ensemble methods too. The reasons behind this are manifold.

Firstly, deep learning models are large, including multiple layers of hidden neurons. This constitutes a challenge to interpreting their functioning. In the best-case scenario, humans may understand the workings of individual layers. However, they have a limited capacity to hold information in memory. Hence, human intuition fails at high dimensions, with increasing number of layers. Moreover, understanding the individual constituents or layers of a model may not be enough to understand the model in whole. This effect is commonly referred to as "more is different"[6]: the whole includes additional, emerging features that cannot be understood simply by looking at the separate parts. The aggregated structure, the model, therefore remains opaque.

Secondly, a main characteristic of deep learning models is their high recursiveness: subsequent layers are fed by and connected to each other. A mechanism that adds yet another layer of complexity. This is particularly the case for *transformers*, which allow memory allocation in neurons, and results in individual layers being no longer simple. Another consequence of recursiveness is that the level of abstraction in the learned knowledge representations is no longer diachronic. Traditional models projected the data into states that became increasingly abstract with depth. This is no longer true for highly recursive layers, where information travels back and forth throughout the different layers.

---

2. For a more in-depth discussion on the different forms of opacity in machine learning see [Burrel, 2016].
3. In addition to competitiveness, internals of models may also be kept secret in the name of security, as discussed by [Sandvig et al., 2014].
4. Even when one possesses the appropriate skills, reading through code, commercial or otherwise, can still be hard in the absence of well-defined standards and practices [Matteas et al., 2005].
5. Importantly, this is a form of opacity that is unrelated to technical skill, and which affects designers, coders and users indistinctively [Seaver, 2013].
6. This is a well-known effect in complex system studies. It was first described by [Anderson, 1974].

Thirdly, with the advent of hybrid and deep ensemble models that combine different architectures comes the additional issue of how these architectures are combined into a unique system. Traditional ensemble models combined outcomes from multiple or gradient boosted tree learners. Understanding them required inspecting the individual learners first and then focusing on the way their outcomes were aggregated. Hybrid models today combine different forms of deep learning architectures, including recurrent neural networks (RNN), long-short memory networks (LSTM), convolutional neural networks (CNN) or generative adversarial networks (GAN) with other model forms, such as support vector machines (SVM) or Bayesian models. Aggregating multiple tree learners is challenging in terms of understanding. Making sense of systems that combine these architectures quickly become unfeasible.

Finally, even when the models themselves are simple, the resulting predictive systems can still be opaque. As will be discussed later in greater depth, machine learning training and delivery is a convolute process that includes several steps. During the data pre-processing step raw variables are often combined to obtain a reduced set of highly predictive attributes that capture the nonlinearities in the data. This process effectively obfuscates the original attributes and results in features that are hardly understandable for the ordinary citizen. Hence, models based on these features, albeit being simple in nature, should still be considered opaque.

Opacity can therefore arise from the size of the models, which can include multiple layers, or from the patterns emerging from the intricate relations among these layers, as well as from the aggregation of several multi-layered structures.

Independently of its source, **opacity and therefore complexity represents a barrier to the large-scale adoption of machine learning**. This barrier can be defined not only as the problems that stem from opacity itself, but also (and perhaps more significantly) as a question on whether complexity brings significant improvements on performance, efficiency, and sustainability.

Figure 1: The challenges stemming from complex, opaque models

Related to performance	Stemming from complexity	Stemming from opacity
→ Not enough evidence supporting a correlation between complexity and opacity	→ Inefficiencies on training and setting up → Unsustainability on resource consumption	→ Business models exposed to blind spots → Users/consumers, especially in high-stakes decision-making → Lack of regulatory compliance

Source: created by the author.

Performance

Much of the current success of complex models stems from **the belief that there exists a *trade-off* between accuracy and interpretability**. The more complex a model is the better it is presumed to perform. This perceived relationship, however, lacks sound experimental proof and remains, as of today, a mere perception[7]. On the contrary, the perceived trade-off between accuracy and interpretability is often reversed in real scenarios[8].

7. One of the first articles to denounce the lack of evidence to support this claim was [Rudin, 2019].
8. An increasing number of voices warn against the fact that there is little to no experimental evidence of the accuracy/interpretability tradeoff from an end user perspective. As discussed by [Herm et al, 2021], this tradeoff is highly situational and dependent, among others, on the considered application and the training data.

In many relevant applications no significant performance differences are observed between complex models and much simpler one[9]. Evidence shows, for example, that simple models are not any less accurate than complex, black-box solutions when evaluated *in the wild* in fields such as the healthcare and criminal justice systems, or even computer vision[10]. Complex models, it seems, might not always be the best option when it comes to developing automated tools to aid the decision-making process in high-stakes contexts, after all.

9. Even so, this is a belief that is deeply engrained in the machine learning community, and which has shaped the research and development agenda of the field during the past decades. To give an illustrative example, hundreds of papers are published every week that improve deep learning performance standards by a few decimals by introducing small refinements on existing methods. In contrast, most decision tree models trained today by both the academy and the industry are based on CART, an algorithm which dates back to 1984. Other algorithms have been proposed since then, including ID3, MARS or CHAID. Yet, CART remains the overall standard when it comes to training decision trees. See [Breiman et al., 1984] for the original publication. Assuming there exists a gap in performance between more complex and simpler models, little progress has been made towards closing it.

10. [Angelino et al., 2018] show that their simple, rule-based models is as effective as the black model COMPAS when it comes to predicting rearrest in the US justice system. [Caruana et al., 2015] discuss two case studies where intelligible models yield state-of-the-art accuracy for pneumonia risk prediction hospital 30-day readmission.

In focus

Additional reasons

The skewed perception on performance of complex models can be attributed to current practices in research and development of machine learning.

- **Improvements attributed to complex models are generally based on comparisons on static data**. This practice disregards relevant aspects of real problems. The process of extracting knowledge from data may require several iterations. It is rarely based on a unique, static evaluation. Moreover, experiments rarely report performance within the extended context where the models will operate in real-life, forgetting critical aspects of how machine learning is served in most companies. More generally, the optimal cost/benefit trade-offs of a problem are usually subject to change and should therefore not be assumed to be static. Hence, we may have collectively overestimated the competitive advantage offered by complex models in real applications.

- **Model performance is often evaluated assuming equal costs**. In many cases models are evaluated by straightforwardly computing the error rate for all samples indistinctively. In adopting this approach, we may be comparing models on scenarios which are substantially different from those where they will be productivized. In fraud identification contexts it is well-known that the cost of false positives cannot be assumed to be equal to that of false negatives. Equivalently, in credit scoring settings the cost of incorrectly estimating the risk of default can be very different depending on the amount of the considered loans.

→ **Trend has a profound impact on how research is conceived and carried out**. Finally, we should not underestimate the impact of trend on scientific publication. Trend today is deep learning. Every week, hundreds of papers are published in top machine learning journals that propose new ever more complex architectures that offer a predictive advantage over existing methods. This advantage is demonstrated by comparing the proposed architectures against simpler ones. The latter are surprisingly always found to perform worst. Whether the efforts dedicated to training those simpler solutions are comparable to those invested in developing the proposed alternatives remains to be seen. Unless they are, gains reported on theoretical grounds may not translate into real progress that supports the use of complex models in real applications.

Complexity

Sustainability. Complex models make an intensive use of computational resources, which can lead to diminishing returns. With the rise of deep learning, models have seen a dramatic increase in the number of parameters. Architectures consisting of thousands or even millions of parameters are commonplace in fields such as natural language processing or computer vision[11]. These architectures can take hours or even weeks to train. Sometimes even more, given that training itself is no guarantee of performance. Machine learning models are trained using a trial-and-error process. Hence, a single training iteration may not suffice to obtain the desired performance. Instead, finding the optimal parameter setting can involve several cycles. Even when the process is over, the obtained solution can be incorrect. Training and troubleshooting complex machine learning models can be a long, tedious, and computationally intensive task. One which can have severe environmental costs. The carbon footprint of black-box models will increasingly be part of the public debate in the years to come.

11. Google's BERT-large and T5-11B models include roughly 350 million and billion parameters, respectively. OpenAI's revolutionary autoregressive language model GPT-3 contains 175 billion parameters. NVIDIA's Megatron-LM requires 8 billion parameters. A more complete overview of the computational cost for most commercial machine learning models can be found at [Schwarz et al, 2018].

Efficiency. Training of transformers, hybrid models or otherwise large, complex nets can incur high costs for companies[12]. Moreover, this is not only limited to training. Given the size of models, performing inference on individual samples can require a lot of additional computation, which adds on top of large training costs. In some cases, such costs come become extreme.

Case in point

AlphaGo

The problem of efficiency is far from new. In the year 2016 Deep Mind launched AlphaGo, a software based on reinforcement learning to play the game of Go. This experiment required 1920 CPUs and 280 GPUs to play a single game, with an estimated cost of $35,000,000[13]. Training increasingly complex models is increasingly unsustainable and costly. More so considering the recent rises in electricity prices. These types of models, now considered a commodity in most sectors, can very quickly become out of reach for many companies. If so, one question that arises is whether the promise of a universal higher performance is reason enough to keep investing exclusively in these models.

12. A single training iteration of BERT-large require use of 64 TPU chips for four days at an estimated cost of $7,000.
13. This model has been the subject of great controversy lately. The original release can be found at [Silver et al, 2016]. In response to public outrage at the impact of training such models, Google has recently released a recipe of good practices aimed at reducing the carbon footprint of artificial intelligence. The full report is still under review. A pre-print can be found at https://arxiv.org/abs/2204.05149. Note that the article fails to ask perhaps the most relevant question: are these models really needed or do any viable alternatives exist?

Opacity

Business models. Opacity by itself poses a risk for companies, who may inadvertently deploy solutions which are flawed[14]. Models may, for example, be found to be biased against certain collectives or minority groups, or be based on wrong assumptions, or lead to inaccurate predictions. Opacity may prevent companies from identifying and resolving such issues before serving their models into production. This can result in a severe financial and/or reputational damage[15].

Users/customers. Opacity also poses a clear risk for users, who may be faced with decisions they cannot comprehend or reason upon[16]. In the worst-case scenario, this may prevent them from vindicating their rights against automated decision making.

Regulatory compliance. The issue explained above has led many governments to promote regulation on artificial decision support tools. As of May 25th,2018, solely automated decision making is strictly prohibited in all member states of the European Union[17]. In cases where individuals are subject to high-stakes decisions based partly in machine learning aid systems, the GDPR states that they are entitled to receiving meaningful information about the logic behind those decisions. Affected fields include (but are not limited to) credit scoring, candidate profiling or fraud identification, all of which have a significant impact on people's lives. Failure to comply with such regulation can lead to substantial financial losses for companies deploying these systems in Europe.

14. Machine learning models can only be as good as the data they were trained with [Crawford, 2013]. When these data are incorrect opacity effectively prevents companies from being able to identify and correct errors.
15. In recent years, many voices have denounced the potential negative consequences of companies and public institutions delegating decision making to models whose inner workings are not fully understood. See, for example, [Eubanks, 2018] and [O'Neil, 2016].
16. As machine learning is being increasingly used to inform high stakes decisions much has been said and written regarding how such practices may affect individuals' right to obtain meaningful information about the mechanism behind automated decision making [Selbst, 2017].
17. Whether an actual right to explanation is recognized by this regulation is out of the scope for this work. I will here align with [Selbst & Powles, 2017] to claim that the regulation can be interpreted to at least recognize users' right to receive the minimum required information to enable them to contest the decisions they are subject to and eventually vindicate their rights against faulty or unfair decisions.

In focus

From low-stakes to high-stakes decision-making.

The use of machine learning first popularized for applications such as advertising, product recommendation or web search. Decisions made in these areas are referred to as low-stakes. They can be a source of revenue for companies, yet do not deeply impact human lives. In contrast, areas such as fraud detection, credit scoring or parole hearing result prediction directly affect people's life. These are referred to as high-stakes decisions. Machine learning first appeared as a tool to aid in the decision-making process for low-stakes applications and has gradually increased its presence in high-stakes contexts, where deep learning has become a standard in many industries. However, low-stakes and high-stakes contexts have different needs and require different practices. In providing no information about how the input features are combined to make predictions, opaque models prevent humans from understanding how individual predictions are made. This lack of understanding, which may be acceptable in low-stakes decisions, can have severe consequences in high-stakes decision-making, insofar it deters users from contesting decisions that have a direct impact on their everyday lives.

03

Is explainability a solution?

The last decade has seen a surge in research aimed at extracting post hoc explanations that may provide such meaningful information by speaking, at least partially, of a trained model's learned logic. This approach takes at its starting point the existence of a complex model that needs explaining. Attempts at explaining how this model works usually consist in replicating its behaviour either locally or globally using another, separate model that may be easier to understand[18]. This approach can bring us closer to unveiling the inner mechanisms of complex models. However, it only offers a partial, incomplete solution that cannot be regarded as final.

Explainability will not improve performance. Techniques oriented to making complex models post-hoc explainable can help unveil their learned logic. Yet, they cannot increase their performance. Explainable methods can describe a model, not modify it. In this sense, explainability does not solve the above described lack of a link between complexity and performance.

Inaccuracy. Explanations are approximations to that which they aim to describe. As such, they must incur in a certain loss of information. Explanations extracted from complex models must be wrong to some extent. If explanations could replicate the functioning of a model to a perfect fidelity, the model itself would no longer be needed. At best, explanations can aim at recovering most of the model's logic. And even in those cases when they do, they cannot be complete: an explanatory method with 95% fidelity is wrong 5% of the times. Post-hoc explanations can therefore never be completely faithful to the models they aim to explain[19].

Relevance. In addition, there's also the issue of how relevant post-hoc explanations are. A method aimed at explaining how a given system works should naturally identify the factors that it considers when making decisions. It should also describe how such factors are used and what their individual contribution is to the outcome. Current forms of explanations often fail to provide such a description.

18. Examples of such techniques are LIME [Ribeiro et al, 2016] and its more recent version Anchors [Ribeiro et al., 2018], as well as SHAP [Lundberg et al., 2017].
19. Explanations can be inaccurate of valid only for specific parts of the feature space. I refer the reader to [Mittelstadt et al, 2019] for a more in-depth description of this issue and its consequences.

Incompletitude. To be complete, explanations should describe the different elements that come to play when designing a machine learning solution. In most real-life deployment scenarios, machine learning models do not exist in isolation. On the contrary, they are only a small part of the larger structure entailed by an artificial intelligence system[20]. Such a system includes the model itself, but also the training data, the data pre-processing strategy, the production infrastructure, or any third-party software dependencies. Explaining a machine learning system requires understanding how these different elements interact with each other. For example, models often combine outside knowledge with the training data. This knowledge can be related to a company's business strategy: a company can choose to exclude certain cases or associate a higher value to others. Understanding how the model uses this knowledge is relevant to explaining its predictions. However, deep learning methods prevent us from accessing this information. Explanations extracted from them will therefore fail to see the whole picture.

Case in point

Saliency maps

Take the example of saliency maps, which are often used to explain complex deep learning models based on image or video data. Saliency maps highlight those regions of the image the model focuses on. This can be used to determine the degree of importance of the different pixels. However, saliency maps tell us nothing about how these pixels are being used by the model. They can identify what the model is looking at but don't provide any insights on how the model is processing what it sees. In the absence of this information, explanations based on techniques such as saliency maps can hardly be relevant, or complete.

20. Relevant actors of the machine learning community have made a strong push towards using the more general system when referring to machine learning with the aim of highlighting the many elements that come into play when devising an automated decision-making tool. See, for example, the Montreal Declaration for a Responsible Development of Artificial Intelligence at https://www.montrealdeclaration-responsibleai.com/_files/ugd/ebc3a3_5c89e007e0de440097cef36dcd69c7b0.pdf

04

Interpretable solutions: an alternative approach

An approach that has received less attention is that of creating models that are interpretable in the first place. Instead of explaining them, models could be required to be interpretable *by design*. So-called *white* or transparent boxes refer to **models which are inherently understandable, trained having interpretability in mind**. These can include decision trees or rule-based systems, as explained above. But also, other forms of knowledge representation methods which are better aligned with human understanding. As discussed below, the use of such interpretable models can have several advantages over post-hoc explainability.

Long-term cost effectiveness. Developing interpretable models can be costly. Yet, in most cases the cost of error far exceeds that of dedicating time and resources to developing an interpretable solution. The advantages in this regard are manifold. Interpretable models enable inspection of their internals. This makes the process of locating and fixing errors in time much more agile.

Flexibility. Interpretable models are also better suited to work in different environments. In understanding how they work, they can easily be adapted to new contexts by modifying certain pieces. In contrast, opaque models can be very fragile. When confronted with scenarios other than those they were designed for, they can lead to significant performance losses, even when the differences among the settings are not substantial. This makes building robust prediction systems based on black-boxes difficult. Given the high reputational and economical cost of incorrect or faulty predictions, this makes interpretable models more profitable in the long run. More so, since, in understanding their mechanisms, bits and pieces of interpretable models may be reused to avoid the high costs of retraining.

This shift in how machine learning models are conceived and served in high-stakes contexts such as the above will, however, not be driven solely by regulation. As the use of this technology expands, users will also increasingly demand transparency if they are to trust it. Recent research suggests that interpretability plays a key role in acceptance of machine learning based products and services[21]. That being the case, **adoption of interpretable models will presumably resonate beyond performance and costs**.

21. The role of interpretability in consumer perception is studied in [Shin, 2021] and [Wanner et al, 2020]

In focus

Anticipating and avoiding regulatory risks

Investing in machine learning models that are interpretable by design should also help companies stay ahead of potential trends and anticipate future problems. In April 2021, the European Commission released the AI Act proposal[22], which will probably be enacted into actual enforce regulations in 2023 or beyond. The AI Act is meant to build on top of the GDPR and will be the first attempt to legislate on AI across the globe. It classifies AI into three risk categories: unacceptable risk, high risk and low or minimal risk. Uses that pose a severe threat to the well-being or privacy of users will be understood to pose an unacceptable risk. Applications such as facial recognition in public spaces, social scoring, or subliminal techniques, that fall under this category will be banned.

Machine learning applications that have a substantial impact on people's lives will be considered to be high risk. This may include biometric identification and categorization, employment management, law enforcement, developing of safety components, access to and enjoyment of essential private services and public services and benefits, education and vocational training access, assignment, or assessment, migration asylum, and border control management, and administration of justice and democratic processes, as well as many other high-stakes areas. In all cases, applications falling under the high-risk category will be subject to strict laws and prohibitions. Among others, high-risk AI systems will be required to be transparent and to allow users to interpret their outputs and use them appropriately. Companies will need to adapt their technology to comply with these new rules if they are to stay competitive in the European AI market in the near future.

22. The full proposal can be accessed at https://eur-lex.europa.eu/legal-content/EN ALL/?uri=celex:52021PC0206

05

How to pursue interpretability in practice

Pursuing interpretability by design in practice will require advances in training and design of models such as decision trees or rule-based systems. Such advances should include research on optimization criteria and pruning strategies for trees, as well as development of new techniques for binary and multinary rule design. Other alternatives, such as linear models, including logistic regression, could also be of interest. Independently of their specific structure, interpretable models should be restricted in size or, when they are not, should provide a clear segmentation of cases. They should be based in raw attributes, which remain intelligible to a broad audience. An important challenge in this regard will be designing techniques to capture non-linear effects without hindering the overall understandability.

This is not to say that black-box models should no longer be used. On the contrary, there exist multiple areas of application, including those related to text and image processing, where these models offer a qualitative advantage. In those where they presumably don't, exploring new methods and techniques to train inherently interpretable models can offer substantial opportunities.

It is probably in **those areas labelled as high risk**, such as employment, access to essential private and public services, including credit and insurance, education, nuclear physics, or the justice system **where interpretable models will bring greater benefits** in ensuring compliance with the upcoming regulation. These benefits will be extensive to high stakes decisions in other sectors too, especially in those cases where users can directly interact with model outcomes.

In low stakes settings, the advantages of complex models will probably outweigh the disadvantages. These include optimization of production processes, online advertising, or product recommendation where decisions made by machine learning systems don't have a direct impact on people's lives. In such cases, model understandability may not be an imperative, or incomplete explanations may suffice to satisfy transparency needs.

In this regard, **a distinction should be made between high and low-stakes settings from the users' and the companies' perspective**. Decisions made to advertise one product over another or to recommend investment on one service over another can be crucial for a company, despite not being high stakes from the user side. In such cases, the regulatory framework will play no role in choosing the right modelling approach.

Yet, companies may wish to have a clear understanding of why those decisions are being made, nonetheless.

06

Conclusions

The rise of ever more complex deep learning models poses a challenge in many applications in terms of complexity and opacity. One that has not been yet fully resolved. Unlocking this challenge taking a clear path forward now will both help create value and avoid unforeseen costs in the medium term.

The use of complex models arises from a deeply engrained belief that most problems are inherently complicated and that therefore intricate models are required to solve them. Many, probably most, of these problems are indeed complicated. Yet, **complicated problems need not necessarily require complicated solutions**[23]. Often, the complexity lies in the search for a simple solution.

Following the arguments exposed above, it seems conductive towards efficiency and performance to demand that **the choice of a black-box model over an interpretable solution to be based on opportunity rather than on pre-conception**. Interpretable by design models may not suffice to fulfil the performance or otherwise usage requirements of certain applications. In such cases, black-box models may offer a competitive advantage, provided claims made in this regard are sufficiently backed by solid empirical evidence. In all the other cases, investment in simple solutions should be considered first.

Moving towards a scenario where interpretable models are given a higher centrality may come at a **transitional price**. Doing so in time, however, will present private stakeholders with a **unique opportunity to prepare themselves for a future scenario** where costs of training and evaluation will probably increase, where new advances in regulation will demand that high-stakes decisions be transparent, and where users will not trust products and services that they cannot sufficiently comprehend. Leaving the current stage behind in the short term will contribute to a mass adoption of machine learning and therefore will have a positive impact in the medium to longer-term.

23. One could make an *Occam's-razor*-style argument here to claim that simple solutions may exist to solve complicated problems in many domains. Some researchers have made similar arguments. See, for example, [Rudin, 2019] who introduces the *Rashomon* set argument, or [Hand, 2006] who discussed the *flat minima* notation to show that simple, yet accurate solutions must exist in most cases.

a1

Annex 1 | Challenges ahead for transparent box models and how to overcome them

Figure 2: Challenges for transparent box models and how to address them

Challenge	Stemming from complexity
Lack of skills & tools to developing simpler solutions	Reorientation from technical competence towards analytical thinking + Investment on basic science + Investment on applied science + Adaptation of candidate selection strategies towards new goal
Loss of assurance of secrecy over proprietary solutions	Emphasis on medium-term competitive advantage from companies thanks to → Users' preference towards interpretability → Performance & efficiency gains → Minimization of errors

The paradigm shift towards more interpretable models is not without its own challenges. Interpretable models can entail significant effort in terms of domain expertise. AI education and training today focuses primarily on methods. A theoretical understanding of problems and their implications is often left aside in favour of action. Skills oriented to developing simpler solutions that pose a theoretical challenge are therefore scarce. The tools available to develop such solutions are also limited, and often outdated. If these tools are to be adopted, efforts should be directed towards **developing largely accessible software to use them and bringing them** back to the centre of debate. As happens with so many areas of science, knowledge about interpretable systems will expand as does the community dedicated to their study and practical implementation.

For this to happen **investment should be directed to basic science**. Having a better understanding of what the process of learning entails will positively revert in how models are designed. For example, a common belief is that black-box models can better identify relevant patterns in the data. However, if such patterns are indeed relevant it's possible that, with the right techniques, an interpretable model may also be able to find them.

Moreover, **investment should also be oriented towards applied science**. For researchers to design interpretable models, the technology must exist to do so. This can be a challenge in terms of recruiting. Candidate selection strategies may have to readjust based on the new set of skills required. Most professionals working in artificial intelligence today are well versed on software such as *Tensorflow* or *Pytorch*, which enables training of deep learning architectures. As mentioned above, dealing with interpretable models may require a completely **different set of skills**, less focused on technical competence and more **oriented towards analytical thinking.**

On top of this, there's the issue of **how to ensure secrecy and rights over proprietary solutions if those solutions are to be interpretable**. Many companies today make profits from the intellectual property afforded to black-box models. If interpretable models were to be preferred, those profits would be obliterated, therefore leaving companies to adapt their business models to the new scenario[24]. This can be an issue for companies striving to build their own market share in the machine learning sector. If the internals of models used to inform high-stakes decision-making are to be made public and accessible, this may prevent companies from making legitimate profit out of them. It therefore must be stressed that **companies will still gain competitive advantage** by developing interpretable models which are better perceived by users, or which yield better results than existing ones, and that this will presumably have a positive impact on their balance sheet.

24. Today, an expression of an algorithm in a source code file or programming script can be copyrighted. The algorithms themselves cannot. Companies can prevent unauthorized reproduction of their source code or protect specific products and services based on that code, but the algorithms behind that code cannot be patented.

References

Anderson, W.P. (1972). *More is Different*. Science, 177, 4047, 393-396

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M. & Rubin, C. (2018). *Learning Certifiably Optimal Rule Lists for Categorical Data. Journal of Machine Learning Research*, 18, 1-78.

Article 29 working party. (2017). *Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679*. wp251, Available at https://ec.europa.eu/newsroom/article29/items/622227/en [3.10.2022].

Breiman, L., Friedman, H. J., Olshen, A. R. & Stone, J. C. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, California.

Burrel, J. (2016). *How the machine 'thinks': Understanding opacity in machine learning algorithms*. Big Data & Society, 3, 1-12.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2015). *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 14-16 August, Sydney, Australia.

Crain, M. (2018). *The limits of transparency: data brokers and commodification*. New Media & Society, 20, 1, 88-104.

Crawford, K. (2013). *The hidden biases in big data. Harvard Business Review*. Available at https://hbr.org/2013/04/the-hidden-biases-in-big-data, accessed 7 July 2022.

Dawes, M. R., Faust, D. & Meehl, E. P. (1989). *Clinical versus actuarial judgment*. Science, 243, 4899, 1668-1674.

Domingos, P. (2012). *A few useful things to know about machine learning*. Communications of the ACM, 55, 10, 78-87.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police and punish the poor. St. Martin's Press, Nueva York.

European Union: European Commission. (2021). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing an EU common list of safe countries of origin for the purposes of Directive 2013/32/EU of the European Parliament and of the Council on common procedures for granting and withdrawing international protection, and amending Directive 2013/32/EU, COM/2021/206* final. Available at https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206 [10.08.2022]

Hand, J. D. (2006). *Classifier technology and the illusion of progress*. Statistical Science, 21, 1, 1-14.

References

Herm, L., Heinrich, K., Wanner, J. & Janiesch, C. (2021). *Stop ordering machine learning algorithms by their explainability! A user-cantered investigation of performance and explainability*. In proceedings of the 20th Conference on e-Business, e-Services and e-Society (I3E), 1-4 September, Galway, Ireland.

Lundberg, M. S. & Lee, S. (2017). *A unified approach to interpreting model predictions*. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 4-9 December, Long Beach, California, USA.

Matteas, M. & Montfort, N. (2005). *A box, darkly: Obfuscation, weird languages, and code aesthetics*. In Proceedings of the 6th Annual Digital Arts and Culture Conference, 1-3 December, Copenhagen, Denmark.

Miller, A. G. (1956). *The magical number seven, plus or minus two: some limits on our capacity for processing information*. Psychological review, 63, 2, 81-97.

Mittelsadt, B., Russell, C. & Wachter, S. (2019). *Explaining explanations in AI*. In Proceedings of the Conference on Fairness, Accountability and Transparency, 29-31 January, Atlanta, Georgia, USA.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*, Broadway Books, Nueva York.

Official Journal of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Patterson, D., et al. (2020). *The carbon footprint of machine learning training will plateau, then shrink*. TechRxiv, Preprint, Available https://doi.org/10.36227/techrxiv.19139645.v4, accessed 9 October 2022.

Provost, F. & Fawcett, T. (2001). *Robust classification for imprecise environments*. Machine Learning, 42, 203-231.

Ribeiro, T. M., Singh, S. & Guestrin, C. (2016). *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, 13-17 August, San Francisco, California, USA.

Ribeiro, T. M., Singh, S. & Guestrin, C. (2018). *Anchors: High-Precision Model-Agnostic Explanations*. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2-7 February, New Orleans, Louisiana, USA.

Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence, 1, 206-215.

References

Rudin, C. & Radin, J. (2019). *Why are we using Black box models when we don't need to? A lesson from an explainable AI competition*. Harvard Data Science Review, 1, 2. Available at https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8. [09.09.2022]

Sandvig, C. (2014). *Seeing the sort: The aesthetic and industrial defence of the algorithm*. Journal of the New Media Caucus, 3, 3, 1-21.

Seaver, N. (2014). *Knowing Algorithms*. Presented at Media in Transition 8, 3-5 May, Cambridge, Massachusetts, USA. Available at https://digitalsts.net/wp-content/uploads/2019/03/26_Knowing-Algorithms.pdf [07.07.2022]

Schwarz, R., Dodge, J., Smith, A. N. & Etzioni, O. (2020). *Green AI*. Communications of the ACM, 63, 13, 54-63.

Selbst, D. A. & Powles, J. (2017). *Meaningful Information and the Right to Explanation*. International Data Privacy Law, 7, 4, 233–242.

Shin, D. (2021). *The effects of explainability and causability on perception, trust and acceptance: Implications for explainable AI*. International Journal of Human-Computer Studies, 146, 102551.

Silver, D., et al. (2016). *Mastering the game of Go with deep neural networks and tree search*. Nature, 529, 7587, 484-.

University of Montréal. (2018). *Montréal declaration for a responsible development of artificial intelligence*. https://www.montrealdeclaration-responsibleai.com/_files/ugd/ebc3a3_5c89e007e0de440097cef36dcd69c7b0.pdf [23.10.2018]

Wanner, J., Herm, L., Heinreich, K. & Janiesch, C. (2020). *A social evaluation of perceived goodness of explainability in machine learning*. Journal of Business Analytics, 5, 1, 29-50.

esade

Santander X Innovation Xperts

www.santander.com/santander-x-innovation-xperts-en

By Santander