Deloitte.

Transparent & explainable AI to enhance the effectiveness of name screening

MAKING AN IMPACT THAT MATTERS since 1845

Intro 3
Potential of AI in name screening 4
Lack of scalability in current name screening methods 4
AI and ML as an opportunity for more scalable name screening 4
Case study 1: Advanced analytics in name screening 6
Key requirements for AI in AFC: model transparency and explainability 7
Regulatory background 7
Model transparency 7
Model explainability 7
Case Study 2: Model transparency and explainability in AI-enhanced name screening 9
Case Study 3: Regulatory assessment for AI systems in name screening 10
Conclusion 11
Excurse: Regulation of AI in AFC 12
EU AI Act 12
BaFin principles on AI 13
Wolfsberg Principles of AI 14
If you want to know more 15

Intro

Sanctions are becoming an increasingly more complex topic, due in no small part to the tensions in today's geopolitical climate. Therefore, an ineffective name screening program can create unnecessary risks for any financial institution (e.g., regulatory fines or reputational risks). The rise of geopolitical risks, and the sanctions that come with them, have added more complexity to the already complex task of name screening and put a lot of pressure on back-office operations. Every day, financial institutions have to screen countless number of customers, suppliers, counterparties, etc. around the world, which can be a tremendous task given the wide range of data sources, languages and other factors at play—not to mention the manpower required to manually review the massive number of names flagged and to distinguish between false positives and true matches.

Advanced analytics, and in this context artificial intelligence (AI) and machine learning (ML), offers a promising method to automate much of this work upfront. However, there are risks and a lot of regulatory scrutiny associated with the use of AI, especially in sensitive areas such as anti-financial crime (AFC), where regulators have set the bar very high when it comes to developing, deploying and utilizing AI models. In this white paper, we highlight the potential of AI to make name screening more efficient and effective, but also the regulatory requirements for the safe use of AI in the AFC space. We provide insight into practical, regulatory-compliant ways to use AI-enhanced name screening tools developed by software providers specialized in AFC.

Potential of AI in name screening

Lack of scalability in current name screening methods

Financial institutions are obliged to screen their customers and counterparties names against sanctions lists and other watchlists during the onboarding process and periodically throughout the relationship. In that sense, name screening is an essential part of the know-your-customer (KYC) and anti-money-laundering (AML) procedures that companies in the financial sector employ to meet regulatory and legal requirements. The name screening process also helps firms to get a better understanding of the origin and business activities of their customers as well as any financial crime risks that exist.

Name screening can be defined as a technical process that compares customer data (of natural and legal persons) to public and internal watchlists of sanctioned entities, politically exposed persons (PEPs) and negative news stories.

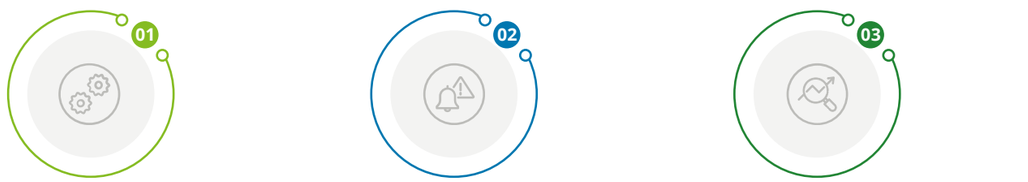
The matching of names and other biographical attributes like country of birth or date of birth, however, cannot be exact, as this would cause firms to miss genuinely sanctioned individuals & entities due to simple misspellings or transliteration errors. That is why it is so important to cast a relatively wide net and capture a large number of false positives along with the true matches. Fuzzy logic algorithms can be effective at reducing false positives by making decisions based on degrees of truth rather than a binary true or false, which will make the screening system more efficient and accurate.

Another frequent source of false positives is homonyms, which are particularly common in certain regions. For instance, one financial institution found that the quality of its screening results was lower in China than in other territories, with the majority of alerts triggered by only a handful of extremely common names. The impact of homonyms may be less dramatic in other countries, though the share of true matches in the total count never goes much higher than a few percent.

Historically, financial institutions have had no choice but to rely on human judgement to assess the names flagged by their screening systems. This is clearly not a scalable approach, as business growth—or even just stricter regulations—will require more and more manpower to process alerts. Fortunately, we now have a number of AI and ML tools that companies can use to reduce the manual effort involved in the process.

AI and ML as an opportunity for more scalable name screening

Advanced analytics helps financial institutions build a more effective and more efficient sanctions screening process in three key ways:



Improved matching

AI can improve the matching process itself, by making it easier to compare attributes.

Alert triage

AI-driven tools help prioritize/triage flagged names. This is the area we have seen the most benefit for financial institutions.

Detection of sanctions circumvention

Network analysis is one way financial institutions can quickly identify the owners and ultimate beneficial owners (UBOs) of corporate clients.

Improved matching

Before you can compare customer attributes with watchlist records, the data has to be uniform and standardized: this requires e.g. transliteration of names to a common alphabet, translation of occupation titles or parsing of addresses and dates of birth. AI has the potential to streamline—or even revolutionize—all of these tasks. Large language models (LLMs) can identify semantic similarity in words and sentences, a huge asset when matching free-text fields such as occupation. Addresses are also traditionally difficult attributes to parse, as they are written differently depending on the location ("Should the system flag someone living on Rue de Téhéran in Paris?"). Machine learning models trained on open-source geographic data can be extremely useful in this task.[1]

Alert triage

Financial institutions typically have large volumes of historical matches that have been manually reviewed in the traditional alert clearing processes currently used by most regulated companies. This historical data is ideal for training supervised machine learning models.

Supervised machine learning models use two different data elements: features and labels. Features, in this case, are numerical metrics that reflect the similarity between the customer attributes and those of entities on the watchlist, for instance the Levenshtein distance between two last names or the number of days separating the dates of birth. In other words, these are the same factors a human analyst would evaluate, but in a language that the algorithm can understand. Labels, on the other hand, are the target of the model, i.e., a measure of what the model is trying to accomplish—in this case, distinguishing between false positives and true matches. Given a sufficient number of examples of both types of matches, these models, assuming they are properly trained, can learn how to use the structure and patterns provided in the features to correctly represent the target, i.e., to identify which of the flagged entities are indeed on the watchlist. The same models can then process any new alerts and classify suspected false positives as a lower priority or automatically close the case, if permitted by law. As the majority of sanctions alerts are false positives, correctly identifying even a fraction of them could save financial institutions a lot of time and money.

There are, of course, a number of precautions to take if you intend to automate alert processing:

1) First, it is important to accurately validate machine learning models to make sure false negatives are not likely to be automatically closed. In other words, the goal is not to reduce the workload of financial institutions by increasing their risk exposure.

2) Second, to the extent that auto-closure is permitted by law, companies need to continually spot check and manually verify auto-close alerts to assess whether the models are producing the intended results.

And on that second point, even if the models have been calibrated to avoid false negatives, there is no guarantee that live production data will be identical to training/validation data. The performance of machine learning models is, in fact, expected to degrade, or "drift", over time as production data changes, even only slightly. Spot checks enable financial institutions to identify those changes as they occur, while also retaining a high level of confidence in the model. And in that process, the model generates new labeled datasets that can be used for the next training iteration of the machine learning model, once performance of the current model has degraded beyond a certain point.

Detection of sanctions circumvention

When onboarding corporate customers, financial institutions not only need to screen the entities themselves but also all individuals and entities that are part of the corporate structure. This screening process can be very complex, with several ownership layers shielding the actual UBOs, who may be many degrees of separation away from the original customer.

It is the job of the analysts to recreate this structure, an extremely time-consuming task that brings together data from multiple, distinct source systems to make a mind map of the links between various entities and individuals. Using graph and network analysis, financial institutions can merge numerous internal and external sources of data to obtain a comprehensive view of corporate structures in a matter of seconds. Having a comprehensive overview helps financial institutions identify potential loopholes or irregularities and take a proactive approach to mitigating risks related sanctions circumvention.

In summary, companies in the financial sector can use advanced analytics to improve the sanctions matching process, automatically close-out the most obvious false positives and quickly escalate any matches that remain for in-depth review.

1 The Levenshtein distance is a numerical value that aims to measure the similarity between two strings. The more significant the difference between the two strings, the higher the number.

Case study 1: Advanced analytics in name screening

In this first case study, a leading technology firm helped a major insurance company based in Central Europe to roll out an effective machine learning-enhanced triage solution for name screening matches. The insurance company, like many other in the financial sector, produced a lot of false positives with its traditional screening system and had to subcontract the manual processing of these alerts to a large number of offshore analysts, which was not a cost-effective solution.

The tech firm supplied a name screening tool with several machine learning models trained on historical data (both alert features and disposition targets) with attributes specific to the regions in which the insurance company operated. The client found that it could automatically close-out a substantial number of false positives thanks to the machine learning scoring, without increasing the risk of accidentally closing a true match (i.e., ensuring no false negatives occur). During the test phase, the share of auto-closures ranged from 30 to 60% of the total alert volume, depending on the respective region.

Combining this kind of automation with the gains in manual investigation efficiency resulting from Machine Learning model explanation and the usage of technology such as graph analytics, it was estimated that the cost savings would be more than 80% compared to the initial situation. Based on this estimate, these savings would relatively quickly offset the licensing fees and support costs associated with the tech company's name screening solution.

Key requirements for AI in AFC: model transparency and explainability

Regulatory background

Regulators around the world are issuing guidance documents and regulations to ensure the safe use of AI and ML. Notably, the European Council and the European Parliament reached a preliminary agreement on the "AI Act" for risk-based regulation of AI and ML systems, while supervisory authorities such as BaFin in Germany have published guidelines on the use of AI and ML in the financial sector.

In the following section, we will examine three key resources, the AI Act, BaFin's guidelines and the Wolfsberg Group's principles on AI. Each framework is unique in its own way, though there is considerable overlap with regard to accountability, transparency and risk management as well as an insistence on adequate human input in the decision-making process. All three emphasize the need for model transparency and explainability when it comes to complying with AI regulations.

Given the regulatory requirements and the expectations of supervisory authorities—especially from a solution provider's perspective—it is clear that the successful use of AI and ML in the broader context of anti-financial crime, and therefore also in the area of name screening, depends on the transparency and explainability of the models in use. When it comes to deploying AI and ML, two main concepts are at issue:

1) Overall transparency and traceability in the development and the operation of AI, including factors from documenting test runs to continuous monitoring.

2) Explainability of the model at the global and local level to provide various model stakeholders with human-consumable explanations of all necessary, decision-relevant information.

In the final section of this white paper, we offer an in-depth discussion of and further insights into the current regulatory requirements, including the EU AI Act, BaFin Principles on AI, and the Wolfsberg Principles for Using Artificial Intelligence and Machine Learning in Financial Crime Compliance.

Model transparency

Due to differences in customer populations, the sets of attributes that are collected and the structure of names as well as alphabets, it is not possible to use the same model across jurisdictions, or sometimes even across lines of business, when developing machine learning solutions designed to optimize the sanctions screening process. That is why it is so vital for firms to develop and train different models for each specific jurisdiction. They may also choose to use different algorithms or different sets of features on different datasets in one particular jurisdiction, simply because one works better than another on a certain dataset. As a result, we can expect a wide variety of models and model versions in this process.

This is even more pronounced in the traditional AI development process, which typically requires some trial and error to figure out what works best. You need a robust framework to manage these models and continually track the data selection and transformation process, the decisions regarding which features to include in a model, what algorithm to use and how to assess model performance, as well as what version of which model (developed by whom) to deploy in production. Companies need to document all of the decision-making here and make it easily accessible in the event of a regulatory audit.

It is well known that most data science solutions for business do not fail because of how difficult it is to train a model, but because of how difficult it is to deploy the solution correctly, integrate it into existing business processes and ensure that business users can correctly interpret the results. To solve this last challenge, you not only need the model management and governance practices we discussed above, but also an easy way to understand how the machine learning models work and make predictions.

Model explainability

Model explainability, or explainable AI, comes in two flavors: ***global*** explainability and ***local*** explainability. The former is measured during model training, communicating the most important features in the new model and the way the model output/decision reacts on average to changes in those features. It helps modelers get an idea of how a model comes up with its predictions.

Local explainability, on the other hand, relates to the individual decisions of the model. It quantifies the degree to which each feature impacts a particular prediction. Although it should correspond to the global explanation on average, there may be significant fluctuations from prediction to prediction. Local explainability can be used, at least to some extent, to justify model decisions on individual alerts.

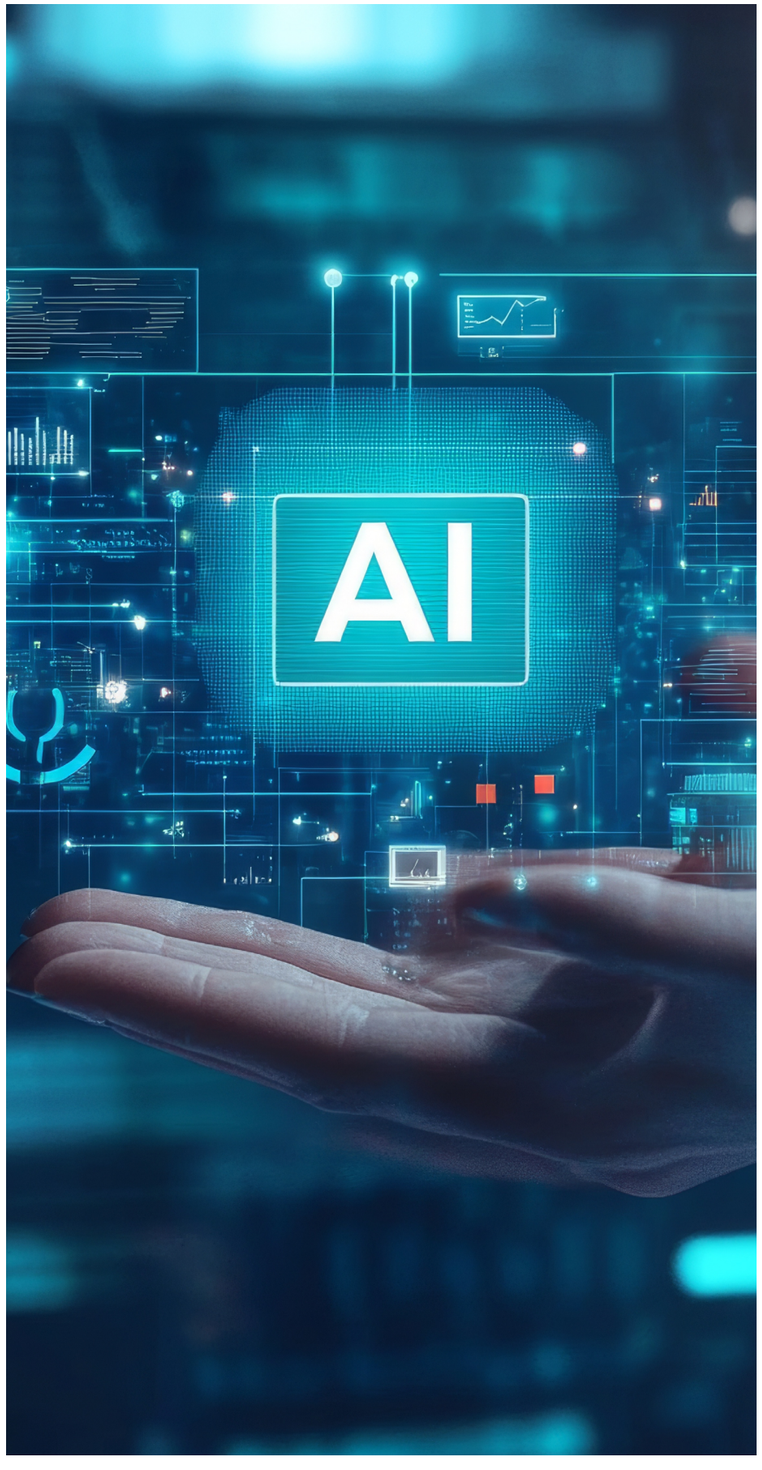
Both global and local explainability are important if you want to provide a comprehensive overview of how and why an AI solution works. However, we should note that a fully deterministic explanation of the inner workings of a machine learning model is usually not possible. In fact, most models do not look at features individually to calculate a prediction, rather, features interact with each other in complex and nonlinear ways, that can only be partially captured by model explainability functions.

One exception here would be simple algorithms such as logistic regression algorithms, which are more easily explained. In most cases they provide less accurate results than their more complex counterparts (such as gradient boosting machines and random forests), but under certain circumstances it might make sense to trade model accuracy for greater transparency.

In the context of explainable AI, it is fundamental that business users understand how a machine learning model arrives at a decision and that they can explain this to regulators in the event of an audit. The most common way to do this is through a simple visual diagram showing the features based on their importance. These are just visual representations of the weight each feature carries in a particular decision, as produced by local explainability tools.

You can supplement these diagrams with dynamic, pre-packaged narratives that explain these visuals in natural language. This has the advantage of speeding up the investigation process, particularly as the system already generates the majority of an audit report for the analyst, usually requiring only a few edits before submission. Tasks like these are obvious candidates for generative AI, and in fact many companies are starting to use LLMs to create more accurate and comprehensive model explanations and audit reports.

Another relatively recent development is the widespread use of machine learning models to provide probabilistic outputs. Sometimes called “Bayesian models”, these models not only compute an individual prediction or score for each alert, they also provide a confidence level. This way, business users not only obtain a decision, but also an indication of how confident they can be about that decision. As a result, the model is much more transparent, while also enabling regulators and financial institutions to have a higher degree of trust in the model.



Case Study 2: Model transparency and explainability in AI-enhanced name screening

As discussed above, companies use advanced analytics in different ways to make their name screening tools more effective and more efficient. The method with the most benefit for financial institutions is machine learning-enhanced alert triage. Historical data on alert decisions, which most companies have on record, is used to train supervised models, enabling automated closing for a significant share of new alerts—or at least accurate prioritization of the alerts that require or do not require further investigation.

Developing and deploying these models can be very challenging in terms of management and governance. In the case study, the financial institution used a solution with a robust framework for AI model management and governance provided within these advanced analytics solution helps financial institutions with all the above, and more. Specifically, this platform provide isolated environments called sandboxes that are populated with select copies of production data that have been identified for specific purposes. This allows companies to retain constant control of the data used to train the ML models without interfering with the inner workings of the production database.

The AFC solution can also provide pre-packaged tools for feature engineering that have been optimized for alert triage in name screening systems. Companies can also build their own functions, decide what algorithms to use during training and select the right hyperparameter values and validation metrics for their purposes. All of these decisions are logged and fully auditable.

Finally, the AFC name screening solution provides a two-step review and approval process (which is also auditable), along with functionality for easy deployment of the respective algorithms in production. While real-time deployment can be supported, optimizing name screening works best in batch processes. This allows modelers and data scientists to focus on finding the best solution for the problem at hand, without wasting time with versioning and repeated explanations of their decisions.

As previously noted, explainability is a key characteristic of machine learning models. It is generally not feasible to provide a fully deterministic explanation of the inner workings of a model—except for the simplest models—as this would defeat the purpose of using machine learning in the first place. There are, however, ways to offer guidance on how supervised models arrive at a prediction. The AI solutions developed by the AFC software vendor are explainable by default; they not only provide global explainability to support data scientists in the model development process, but also local explainability to make individual model predictions and decisions more transparent. The solutions also generate model explanations in natural language narratives, which is easier for analysts to understand and speeds up the process when further manual review is required.

Case Study 3: Regulatory assessment for AI systems in name screening

In addition to many services around the assessment, implementation and use of AFC technology, Deloitte also supports clients in their regulatory journey related to AI and ML used in AFC. In this context, Deloitte supported a German subsidiary of a large European bank as the staff prepared the necessary process and regulatory documentation for a new ML-based solution to reduce the alert triage workload of their legacy name screening solution.

Like many others, the client was burdened with a high number of alerts generated by the name screening solution, which then had to be cleared by a large team of human analysts. It became clear, as they researched possible solutions, that a full replacement of the current solution would not be feasible in the short run. As an interim solution, they decided to automate the clearing process for first level alerts using machine learning models. We helped the client prepare a regulatory strategy for AI during the pre-go-live phase including advice on governance, process and documentation.

The guidance provided by Deloitte relies on sources such as the most recent version of the AI Act, the BaFin guidelines and the Wolfsberg principles, which we cross-referenced. We used these resources along with others to review the overall process the client had set up for the ML model and to determine whether there was adequate human oversight and responsibility in the process. We also looked at the data strategy in use, the governance model and the method used to validate the independent model, the model monitoring process as well as the contingency policies in place. In particular, the review also focused on model transparency and explainability practices with respect to regulatory compliance.

In terms of model transparency, Deloitte assessed whether the bank properly tracks and documents its test runs, whether the results are sufficiently reproducible and which metrics are used for model monitoring, for example, to detect bias and unfairness. In terms of explainability, we evaluated whether the tools and technology used to explain the model results meet market standards on both the global and the local level and enable the developers and users to adequately perform their tasks. This review, along with recommendations for improvements, enabled the client to regularly engage with regulators prior to going live and receive relevant feedback in a timely manner, ensuring a smooth go-live transition.

Conclusion

AI and ML offer significant benefits in terms of efficiency and accuracy in the name screening process and help financial institutions make better decisions by improving matching, prioritizing alerts and identifying UBOs as well as other entities that may be trying to circumvent sanctions. New technology solutions from AFC vendors show considerable potential for cost reduction and operational efficiency, while also meeting the strict regulatory requirements for model transparency and explainability.

It is also crucial to carefully consider the regulatory guidelines and principles. Institutions such as the European Council, the European Parliament and BaFin in Germany have introduced guidelines to help companies and citizens use AI and ML in a safe and responsible way. The main areas of focus include model transparency and explainability, which promote sound governance, compliance, and trust-building between regulators and financial institutions.

While there are challenges and regulatory scrutiny involved in the use of AI and ML in name screening, focusing on regulatory compliance as well as model transparency and explainability will help financial institutions secure substantial efficiency gains, cost savings and better compliance.

Excurse: Regulation of AI in AFC

Even though there are promising applications for AI and ML in name screening and other decisioning tools, the risk is that they can produce false results at a scale that has serious consequences for the people affected by the decision—especially when models are incorrectly calibrated or drifting. The name screening process in particular, which relies on highly personal and often culturally significant characteristics to support decisions, means that incorrectly calibrated or drifting models can have severe implications for the individuals concerned, ranging from the closure of bank accounts to criminal investigations. The potential that decision makers may become over-reliant on machine-generated results, as well as the "black-box" nature of AI and ML models themselves, only makes matters worse.

These risks, especially in higher-risk areas such as name screening, have caused concern among the general public and regulatory authorities about keeping AI and ML safe to use and led to calls for more regulation. In quite a pioneering move, the European Council and the European Parliament introduced the AI Act which intends to regulate the development and operation of AI and ML systems using a risk-based approach. The BaFin in Germany along with many other supervisory and regulatory authorities around the world have also published guidelines on how to use AI and ML in the financial sector, which is particularly relevant for name screening. This brief overview outlines these regulatory endeavors as well as providing additional insight into the principles published by the so-called Wolfsberg Group, a non-governmental association of global banks, as it relates to using AI and ML systems to combat financial crime.

In our analysis, we pay close attention to the three specific sources mentioned above. However, it is important to keep in mind that this is only a fraction of the literature that has been published on this subject to date. There is a variety of information available, and we encourage you explore the topic in more depth using sources beyond those cited in this white paper.

EU AI Act

At the time of writing, the AI Act has come into force on the 1st of August 2024 with the first provisions applying for companies already in February 2025. In the AI Act, the definition of "AI Systems" not only refers to the algorithm itself, but also to the way it interacts with its surroundings (e.g., how it impacts downstream decision-making processes). AI systems must therefore be evaluated not only based on the AI model itself, but also on the entire process and its impact on other factors in its environment.

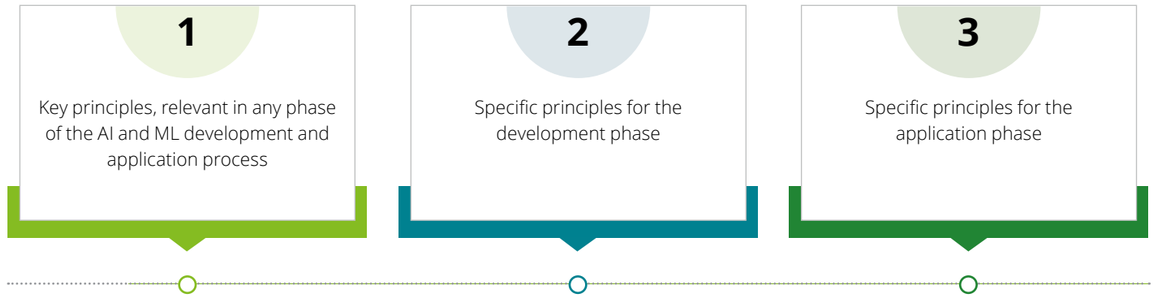
The AI Act uses the risk-based approach common in AFC compliance, allocating all AI systems into risk classes ranging from "unacceptable" to "high" or "low". As a first step, the AI Act requires firms to determine the risk category of the AI systems it is marketing or using. The activities prohibited under the Act are, for example, systems that subtly manipulate, exploit vulnerabilities or so-called social scoring systems, while systems supporting critical infrastructure or law enforcement (that may interfere with people's fundamental rights) are considered high risk. The Act bans the use or application of these prohibited systems, and high-risk systems are subject to certain requirements.

A company must therefore define its own status with regard to the obligations of the Act. The main focus is on the providers of high-risk AI systems and their obligations. Depending on the circumstances, these obligations may extend in part or in full to other obligated parties, such as manufacturers, distributors or users. The obligations for users include, but are not limited to, using the AI system only as instructed, ensuring adequate supervision and monitoring, properly recording the input data, reporting any risks observed, retaining automatically generated audit logs and conducting a data protection impact assessment.

The main obligations for providers of high-risk systems are to set up a risk management system (RMS) designed to identify, analyze and assess risks and to develop risk control policies. Like other compliance-related RMSs, AI RMSs should be seen as a continuous iterative process that requires regular updates. The Act also mandates that these companies establish a data governance system to ensure, among other things, that the datasets used are relevant, representative, error-free and complete, and take appropriate precautions to protect fundamental rights. When creating AI systems, developers must ensure their process are as resilient as possible to errors, faults, attacks, etc., prepare technical documentation and set up an automated record system that logs all operations and events of the AI system. These measures will help ensure, among other things, that the AI system continues to function as designed. Finally, these systems providers are obliged to issue relevant information, e.g., user manuals, to ensure the users can interpret and use the results of the AI system in the appropriate manner. They must also design the AI systems such that human users can monitor them effectively and stop or reset them at any time if they have cause for concern.

BaFin principles on AI

In addition to the AI Act, which applies on a global level and across multiple use-cases, the expectations of financial supervisory authorities have a significant impact on the financial sector. The use of AI and ML will undoubtedly play a major role in the future of financial institutions, which is why various supervisory authorities around the world have begun to formulate or have already issued their guidance regarding the development and deployment of AI and ML applications. The financial supervisory authority of Germany (BaFin), for example, has issued a set of principles for the responsible use of big data and artificial intelligence (BDAI) in the decision-making processes of financial companies. BaFin published these guidelines on June 15, 2021 to help companies in the financial sector mitigate the risks associated with AI and ML. While BaFin acknowledges the benefits of AI and ML applications for businesses and consumers, it also highlights the need to manage the risks associated with them. BaFin's principles fall into the following three groups:



The first of the **key principles** in the guidelines is that senior management should take full responsibility for enterprise-wide strategies, policies and rules relating to the use of algorithmic decision-making processes (**clear management responsibility**). Senior management must also set up a risk management system—including, where necessary, an effective outsourcing management system—adapted for the use of algorithmic decision-making processes (**appropriate risk and outsourcing management**). To promote good business decisions and customer relationships, companies must also ensure their algorithmic decision-making processes produce unbiased results (**preventing bias**). And finally, financial institutions must implement processes to prevent and monitor discriminatory bias in violation of legal (and ethical) standards (**avoiding types of discrimination prohibited by law**).

The **specific principles for the development phase** stipulate that companies use only data of sufficient quality and quantity (**data strategy and data governance**) as well as comply with applicable data protection requirements (**compliance with data protection requirements**), depending on the application and features of an algorithm. The principles for the development phase also call for accurate, robust and reproducible results (**ensuring accurate, robust, and reproducible results**) and proper documentation of the selection, calibration and training process for models as well as of model validation (**documentation to ensure clarity for internal and external parties**). Regarding model validation, the principles stipulate that an independent entity or an individual not involved in the original modeling process should be tasked with performing or at least reviewing the validation process (**appropriate validation processes**). Finally, these principles put a particular focus on selecting a balanced dataset for calibration and validation (**using relevant data for calibration and validation purposes**). Using unbalanced data in the calibration or validation process can lead to modeling bias, and it is important to prevent this bias in as part of the data preparation stage.

The **specific principles for the application phase** stress that, depending on the mission-critical nature and risk level of a decision-making process, employees should be actively and meaningfully involved in the interpretation and use of results generated by algorithms (**putting humans in the loop**). This includes, for instance, the manual approval of results when the algorithmic results exceed certain thresholds, even in otherwise automated processes (**in-depth approval and feedback processes**). In addition, financial institutions must have contingency measures in place to ensure business continuity should any issues arise (**establishing contingency measures**). Finally, companies must regularly validate their algorithms throughout the entire lifecycle of a model in order to assess their functionality and identify any discrepancies based on predefined parameters (**ongoing validation, overall evaluation and appropriate adjustments**). The models must then be adjusted accordingly based on the results of the validation.

Wolfsberg Principles of AI
The Wolfsberg Group is a private initiative involving twelve global banks which aims to establish and operationalize standards for the management of financial crime risks. Motivated by the continued regulatory uncertainty surrounding the use of AI and the practical necessity of using AI to fight financial crime more effectively and efficiently, the Wolfsberg Group published "Wolfsberg Principles for Using Artificial Intelligence and Machine Learning in Financial Crime Compliance".

Based on a wide range of previous publications on data ethics, the Wolfsberg Group identified five guiding principles:

01 **Legitimate purpose**
Financial institutions should develop initiatives to combat financial crimes based on the regulatory requirements and a commitment to protect the integrity of the financial system. FIs must guard against the potential misuse or misrepresentation of data and any bias that could influence the results of the AI/ML application.

02 **Proportionate use**
It is up to the financial institutions themselves to find the right balance between the benefits of AI use and the appropriate management of the risk that may arise from these technologies when they develop and deploy of AI/ML solutions to comply with anti-financial crimes regulations.

03 **Design und technical expertise**
Financial institutions must monitor the technology they rely on and understand the implications, limitations and consequences of its use. They should take care to staff the teams involved in the creation, monitoring and oversight of AI/ML with employees that have the right skillsets and a diverse range of experiences.

04 **Accountability and oversight**
Financial institutions are accountable for their use of AI/ML, including those decisions based on their analysis of AI/ML results. They should train their employees in the appropriate use of these technologies and consider oversight of their design and technical teams.

05 **Openness and transparency**
Financial institutions should report openly and transparently on their use of AI/ML, in line with legal and regulatory requirements, but they should take care to ensure that this transparency does not inadvertently make it easier for the industry to commit financial crimes or violate confidentiality and/or other data protection obligations.

In summary, the Wolfsberg Group recommends in its principles that financial institutions should implement the principles for the use of AI/ML as part of their compliance and risk management activities using a risk-based approach, which may vary depending on the evolving regulatory environment and the specific use of AI/ML by the financial institution to combat financial crime.

Considering all the regulatory expectations outlined above, we can conclude that model transparency and explainability are key concepts for broader AI compliance. After all, the appropriate assessment of risks and rewards, suitable calibration and continuous validation methods, meaningful human involvement in the decision-making process and other aspects all require transparent and explainable models.

If you want to know more

Martin Hirtreiter
Deloitte | Partner
Anti-Financial Crime Advisory
mhirtreiter@deloitte.de

Dr. Robert Schmuck
Deloitte | Director
Anti-Financial Crime Advisory
rschmuck@deloitte.de

Julian Sebastian Koller
Deloitte | Manager
Anti-Financial Crime Advisory
jkoller@deloitte.de

Deloitte.

Issue 12/2024