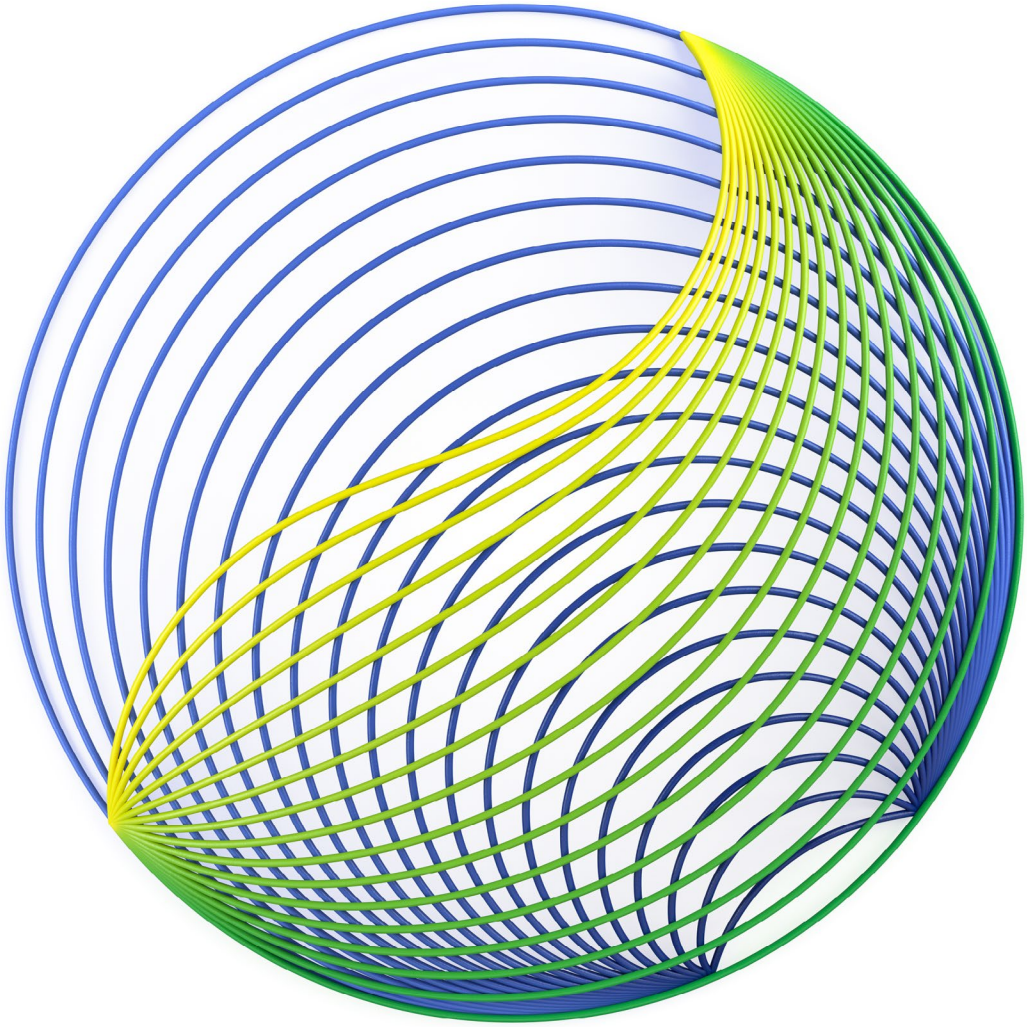
Deloitte.

HAWK



AI in Anti-Financial Crime

Using an Overlay to Leverage AI Without Ripping and Replacing a Legacy Solution

AI is on the rise in anti-financial crime. How can financial institutions manage the transition from legacy systems?
Before AI applications matured enough for use, financial institutions (FIs) installed market-dominant, rule-based monitoring and screening systems to meet regulatory demands. These systems, which were often on-prem solutions, took years to implement, entailed a lot of specific customizations, cost a considerable amount of money and time, and became deeply embedded with the institutions' complex data landscapes.

FIs are now grappling with legacy systems that are often ineffective and inefficient at combating financial crime. As effective AI Anti-Financial Crime (AFC) solutions mature, FIs also want to take advantage of these technological evolutions. However, these often feel daunted by the challenge of replacing their existing systems. AFC professionals in the trenches know how involved and complicated such an implementation is; they have gained extensive experience in these systems for years.

As a solution to this system problem, we propose an "AI overlay" implementation model. This concept involves using AI to enhance the current setup without removing it. For instance, AI could be used on top of—or in parallel with—the results produced by the legacy systems, such as anti-money laundering (AML) alerts or screening hits.

MAKING AN IMPACT THAT MATTERS since 1845

This approach leverages AI to make the current system setup more effective and efficient without extensive overhauls or change processes. The implementation of such an AI overlay is simpler and promises quicker results at much lower costs, compared to a full replacement. Nevertheless, an AI overlay still requires careful consideration of explainability, processing time, and regulatory compliance.

AI overlay keeps the current setup, but enhances it with the Power of AI

What does "AI overlay" mean?

By "AI overlay," we mean leveraging the capabilities of AI without having to change the current system setup. AI is employed either above or alongside the results generated by the legacy system, such as AML alerts. The concept of AI overlay refers to the integration of an AI solution into the existing setup and processes at various stages to enhance the efficiency or effectiveness of these processes.

A case in point is the usage of AI to prioritize and sort rules-driven AML transaction monitoring alerts. In this instance, the AI overlay solution would process the alert data and additional transactional data sourced from the legacy system and apply AI models to compute the likelihood of whether the alerts are true or false positives. The outputs from the AI model alongside other critical information such as AI explainability are then fed back into the legacy system's case management user interface.

The alert handler utilizes this additional information to make quicker and more accurate decisions on rule-based alerts or to prioritize their work in line with the AI results. AI overlay solutions typically learn from historical alert and transactional data procured from the legacy system. They are tuned to anticipate whether a legacy system alert will be escalated or not as part of supervised learning. They can also be trained to identify unseen anomalies within this data set through unsupervised learning. These models are designed by subject matter experts, leveraging a wide range of experience in building models for AFC use cases. Thus, the AI overlay concept represents an optimal solution from both efficiency and effectiveness standpoints.

Why is AI overlay a valuable option?

The concept of "AI overlay" presents a valuable option, particularly when compared to the complete replacement of an existing AFC solution. The main advantages of this approach lie in the AI overlay implementation being significantly quicker and less expensive, while also allowing immediate benefits from using AI. Even though it might not make use of AI at its full potential initially, it provides a competitive edge and time to strategize for a more long-term solution.

The AI overlay's implementation involves deploying single models on already consolidated data from the legacy AFC system, either on-prem, on a private cloud, or via open cloud Software as a Service. The implementation is relatively uncomplicated, requiring less coding based on a common data schema from household name legacy systems. Additionally, the integration into existing processes is smoother, with fewer disruptions. Processes need no extensive overhaul but simply suitable adjustments to accommodate the AI layer integrated into the bank's existing model validation processes.

As alerts' decision-making processes and user interfaces can essentially remain unchanged, employees do not have to endure significant change processes. In essence, AI overlay offers the advantages of AI with lower costs and friction. It provides breathing space for crafting a more thoughtful long-term AI strategy.

Compared to self-built AI overlay solutions, an AI overlay from a vendor model is also a more cost-sensitive solution, as well as a more effective and efficient one. Building models for AFC use cases require data handling proficiency, advanced AFC modelling capabilities, and years of domain experience from data scientists as there are many common pitfalls that need to be avoided. Leveraging this unique specialization in AI and AML allows FIs to save the costs they would incur developing AI technology in-house.

"The main advantages of this approach lie in the AI overlay implementation being significantly quicker and less expensive, while also allowing immediate benefits from using AI."

What are the different types of AI overlay in AFC?

In general, and as a prime use case, an AI overlay is used on top of alerts produced by a legacy system, e.g., rule-based AML transaction monitoring alerts. The model produces results based on data coming directly from (i.e., pre-processed by) the legacy system. This includes the alert data as well as additional data related to the customer and/or transaction being alerted.

Here, the efficiency and effectiveness advantages of the AI overlay come into play. The main drawback of this approach is (most probably) not being able to produce results in real-time. These standard use cases are captured in the upper right corner of the matrix in Fig. 1. Examples of this use case are those where decisioning is not bound to real-time requirements, such as for alert triaging or alert discounting. AML transaction monitoring alerts are usually produced the day after the suspicious transaction occurred (t+1) and are then worked on by the alert investigator. In this use case, a model, provided with additional contextual risk-factors, is trained on prioritizing, or discounting rule-based alerts and feeding this information back into the user interface of the legacy system. This information can then either be used to support subsequent human decision making or even – given regulatory constraints – managing obvious false positives.

Fig. 1 – Potential use-case groups for "AI overlay" solutions

	Source system data	Legacy system data
AI on existing alerts	**n.a.**	**Use case:** alert prioritization, discounting, 1st level review **Real-time:** no **Integration:** relatively simple **prime use-case**
New alert generation	**Use case:** anomaly detection **Real-time:** potentially yes **Integration:** as integrating a new solution	**Use case:** anomaly detection **Real-time:** potentially yes **Integration:** relatively simple

Additionally, an AI overlay can be used not only to prioritize, triage, or discount existing alerts, but also to generate completely new alerts of higher quality (bottom row on Fig. 1). This can also be done by data already being processed by the legacy system (bottom right corner). Here, non-alert data can be used (for example, by an unsupervised model) to identify anomalies otherwise not detected by the rule-based legacy system. This use case complements alert prioritization and builds on the same advantages (easy to implement AI usage). This can be done also in real-time, however if integrated in the existing file exchange-process and connected to a legacy system for transaction monitoring, the AI overlay will not be in real time. Consider AML transaction monitoring anomaly detection as an example. By identifying outliers from regular transactional behavior, the model can identify transactions which would not have been triggered otherwise.

Lastly, new alerts can be generated alongside alerts being generated by the legacy system, but this time, data is retrieved not via the filter of the legacy system, but directly from the source systems or at a compliance data integration layer. In this scenario, for the newly generated alerts, the legacy system is only used for user interface purposes. Strictly speaking, this use case does not constitute an "overlay," as model results are not produced on top of existing alerts, but rather in parallel. The advantage here lies in the fact that a parallel integration of an AI directly to the source may allow – depending on the data infrastructure setup—for faster processing times. On the other hand, however, this approach is comparable to a rip-and-replace integration in scope of work, as all source systems need to be docked onto the new AI layer.

Implementation is comparably simple, with some considerations

Even though implementation of an AI overlay is – compared to a full system rip and replace – generally considered to involve less time and effort, FIs need to consider a few implementation elements.

Explainability

In general, AI models produce risk scores which can be added to the previously generated alerts within the legacy system. However, especially within the field of AFC, each decision needs a proper explanation of why something is considered a risk or not, i.e., each (machine-based) decision needs to be properly explained. This explanatory information is – at best – a break-down of all decisioning factors (i.e., features) used by the model and their respective weighting written in plain, human-readable text. Depending on the number of features used, this can add up to a substantial amount of text data which needs to be forwarded and integrated into the legacy system's case manager. Especially with regards to the latter, FIs need to assess case manager flexibility and implications before engaging in the AI overlay use case.

Processing time

The multi-layer setup (going from source system, to integration layers over the legacy system, to the AI overlay application, and back) reflects itself in potential processing times which need to be considered when evaluating suitable use cases. Especially in an overnight batch processing setup, where the initial transfer from source system to legacy system already implies a t+1 day timespan. Even though the overlay system itself might be lightning fast, an additional batch processing towards the overlay system requires additional processing time requiring clarification from a regulatory point of view. From jurisdiction to jurisdiction, supervisory expectations on processing timespans can vary and need proper consideration before implementing an AI overlay solution.

On-prem vs. cloud deployment

Legacy solutions are often on-prem installations, whereas AI overlay providers usually – as a baseline – deploy on open cloud environments. Even though AI overlays can often be deployed locally, a local installation is probably not compatible with the wider cloud strategy of the FI. Thus, AI overlay is, when not provided as SaaS, deployed at, or interconnected with, different private cloud environments raising multiple compatibility topics when interacting with the on-prem environment where the legacy solution is hosted.

Regulation

FIs need to assure technical compliance elements, such as model validation, model transparency, and model explainability. AI overlays also need to meet the expectations of the banking supervisory authority and get their approval before they can be used in an AFC context. This might imply restrictions on AI use cases. When thinking about an AI overlay use case, a FI should engage with the regulators as soon as possible to discuss their proposed technical and process fulfillment of compliance requirements.

Summary

In this paper, we primarily discussed the difficulties that FIs face with their old, often inefficient AFC systems, and proposed an alternative solution referred to as "AI overlay". Regulatory demands have led to the prevalence of rule-based, on-prem solutions which are intricately linked to complex data. These systems have been the result of large investments and time-consuming integration. Given soaring compliance costs, FIs are looking to leverage AI without incurring heavy costs replacing their legacy setups.

This is where the concept of "AI overlay" comes in. This method allows the current setup to remain while using AI to enhance it. AI systems are used in parallel with AML alerts and transaction monitoring. The AI overlay approach helps increase efficiency, as it offers a faster implementation at lesser cost than a replacement solution. This provides immediate benefits of AI use. Importantly, it also offers a competitive advantage during a period of transition while a more strategic, long-term solution is developed.

We discussed different types of AI overlay, such as prioritizing or discounting existing alerts, identifying unusual behavior, and even running in parallel to create new alerts. We pointed out that the implementation of AI overlay, while simpler than a full system replace, still requires careful consideration in terms of explainability, processing time, and regulatory compliance.

If you want to learn more

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

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

Moritz Schneider
Deloitte | Senior Consultant
Anti-Financial Crime Advisory
mschneider@deloitte.de

Wolfgang Berner
Hawk | Co-Founder & CTO/CPO
wolfgang.berner@hawk.ai

Deloitte.

Issue 08/2024