December 2020

Governance of artificial intelligence in finance

Summary of consultation responses

Author: Laurent Dupont

Fintech-Innovation Hub, ACPR

The 26 written responses to this consultation received by the ACPR originate from banking institutions, technology providers, consultancies, professional associations, and research institutions. In addition to those written responses – most of which answer all submitted questions –, the present summary also reflects a dozen oral conversations about the consultation. Their main benefits are to bolster the principles exposed in the discussion document and to provide guidance for the Fintech-Innovation Hub's future work topics.

Experience and organisation in Artificial Intelligence (AI) and Machine Learning (ML)

AI/ML skills appear well-developed among financial actors' internal staff, and are often enhanced with transversal acculturation and sensitisation initiatives. AI experts are found in dedicated, decentralised structures close to the lines of business, or gathered in a more centralised manner, or even across local entities within a group. Backgrounds vary from academia through applied researchers to experienced practitioners.

Algorithms and use cases

AI algorithms implemented by respondents are more diversified than anticipated: while gradient boosted trees are commonly used, deep learning techniques are less frequent, and unsupervised learning (including clustering techniques) is well represented – as are standard or custom "model-ensembling" methods.

The most frequently cited usage categories are internal productivity tools (e.g. automated document analysis), customer relationship management, specific investment applications, and several processes along the insurance value chain[1].

While the discussion document focused on business-critical processes, a noteworthy point among responses is the number of use cases involving NLP (Natural Language Processing), especially for internal management processes.

Explainability principle

A major, growing concern for explainability emerges from the responses, in relation with performance objectives, regulatory compliance, auditability, and ease of internal adoption of AI.

Respondents deem the 4 explanation levels to constitute an interesting graduation because it is qualitative. They also agree with the proposed definitions, except for the highest (level 4) which is considered unachievable in

[1] For a more detailed review of use cases in the financial sector, see the previous discussion document "Artificial intelligence: challenges for the financial sector" published by the ACPR in 2018.

practice and more related to quality assurance than to explainability. Also, providing an explicit mapping between each explanation level and a set of adequate algorithms for a given use case would help to clarify the boundary between levels 2 and 3.

Respondents are also in line with the two main factors driving the choice of an explanation level, namely the risk associated to the use case and the recipients of the explanation. Additional factors suggested are the application perimeter of the model and the robustness of the explanations produced.

The few tweaks proposed for the practical examples of explanation levels consist in raising the required level for the AML-CFT compliance officer, for those tasked with validating and overseeing regulated models, and for the end consumer (where level 2 or more could be demanded).

The lack of examples involving NLP is again pointed out, although the usefulness of explanations is deemed limited for NLP models insofar as such models can be monitored on the basis of their sole output. Future studies would also benefit from including advanced AI uses in AML-CFT such as graph analysis and unsupervised learning.

Lastly, the articulation of explainability requirements with both regulatory constraints and social issues (e.g. discriminatory practices) should be investigated.

Performance principle

The document enumerated technical performance metrics for AI. This list, provided as a discussion basis, was augmented by respondents so as to cover a broader range of tasks (classification, regression, text analysis) and use cases.

They also emphasize the necessary distinction between metrics intended for data scientists – which aim to optimise the model – and those intended for subject matter experts – which should above all be easy to interpret.

Functional performance metrics can be broken down as normative efficacy metrics (e.g. in financial security: the coverage of identified scenarios or the prevention of the risk of adaptation by fraudsters) and operational efficacy metrics (processing time, quality of raised alerts, risk associated to the alert processing result, etc.).

Stability principe

The most frequently cited sources of model drift pertain to input data, to model retraining, but also to the technical, economical and regulatory environments (new filtering rules, new service offering, new customers, etc.) Associated risks include a performance loss, the introduction of biases, and financial risk.

The proposed mitigation techniques for model drifts are quite classical: continuous monitoring of statistical coherence indicators, preventive detection of generalisation problems, resolution of such problems (via a rigorous sampling procedure, model simplification, transfer learning, non-systematic and less frequent retraining, etc.), and model versioning and reversibility as a stopgap measure.

Appropriate data management principle

Few responses on this point are AI-specific. At an organisational level, a general recommendation is for legal and data protection departments to collaborate with data scientists on the following aspects: compliance of the use case, model purpose, data usage, consequences of the data processing, and implementation of appropriate safeguards.

A tension is also emphasized between regulatory requirements on the data retention period on the one hand, and the necessary storage of decision logs and input data in order to be able to produce individual explanations on the other hand.

Respondents use a range of **bias detection** method: back-testing for validation, explanatory methods, manual review, monitoring of statistical indicators in production. The definition of adequate fairness metrics is identified as requiring clarification by competent authorities.

Respondents seem familiar with the mitigation of **data biases**, less so with the prevention of **model biases**, which is limited to the available features and even deemed superfluous in the case of interpretable predictive models.

Integration in business processes

Respondents note a wide variety of **evaluation criteria for the integration** of AI: internal productivity, customer relationship, human-machine interactions (including explainability and the right to object to algorithmic outcomes), ease of deployment and maintenance, reversibility and error correction techniques, and regulatory compliance. A distinction is proposed between two types of scenarios: those where AI mostly helps to improve accuracy or performance, and those where AI fundamentally impacts the business process, thus inducing a risk of "forced adoption" or even rejection by the business side.

A disagreement regarding **AI autonomy** emerges from the responses: should human agents be encouraged to put algorithmic results in question and thus maintain their freedom of decision, or does this human autonomy instead incur additional risks (due to human overrides of system decisions)?

Parallel processes in which human operators are tasked with continuously assessing the AI's behaviour are not unanimously approved due to their financial and staffing cost and to the operational challenge they represent.

According to respondents, **AI engineering methodology** should evolve in the same general direction as the software industry, specifically toward better quality control and robustness enabled by "MLOps" design principles. An essential difference between AI and traditional software is however underlined: AI (and ML even more so) is built using a trial-and-error approach, whereby different solutions are concurrently implemented in order to gauge their respective efficacy on the decision-making procedure. Their evaluation must therefore be iterative and follow an experimental – albeit rigorous – methodology rather than proceeding by formal proof.

Internal control system

Responses bolster the idea that the integration of AI does not radically modify **business risks** compared to classical, statistical models: whilst operational risk is generally amplified, ML models still fit in governance procedures for model risk management. Furthermore internal control procedures are already supposed to adapt to evolving usage and technology. Additional reputational and legal risks are however pointed out as potential outcomes of discriminatory automated decisions, which raises the question of **liability for AI-driven decision-making processes**.

Methods for mitigating AI-induced risks are in line with the 4 design principles presented in the ACPR's document. Respondents also add standard best practices (model risk mapping, method for reverting to human control) and an increased involvement of the legal department to support and oversee AI projects.

Respondents confirm the importance of **functional validation** (provided it is proportionate to the criticality of business processes involved) across the AI lifecycle, which incidentally requires building AI skills within internal control departments and banking or insurance institution management.

Respondents also underscore the tension between auditability (and thus explainability) requirements for **internal models** in a regulated sector and the promise made by AI to impose minimal assumptions and to automatically adapt to its environment. Views differ on ML applicability to "Basel models": some respondents consider that ML-based risk models can be deployed to production themselves (albeit on a well-defined perimeter, upon demonstrating their benefits, and with guarantees on the 4 design principles), whereas others only envisage indirect uses of ML to improve existing models (e.g. by introducing new features or adjusting business rules).

To the regret of a few respondents, the relationship between **model risk management or MRM** (whose review procedures are often incompatible with a short validation cycle) and the introduction of AI (which carries a human behaviour modification risk usually neglected by MRM) is absent from the discussion document. Responses however confirm the ACPR's argument that the use of ML does not fundamentally call into question an organisation's **internal model update policy**.

Technical validation and continuous monitoring processes are according to respondents quite similar to the case of traditional models. A few attention points are nonetheless underscored: increased complexity of data pre-

processing, bias detection, hyperparameter selection, and monitoring of the 4 principles exposed in the ACPR's document.

Security and outsourcing

Respondents confirm and even broaden the range of **risks related to AI outsourcing**: they include data access issues and data leaks, vendor lock-in, software dependency, lack of information and loss of knowledge, deficiencies in the third-party engineering processes and security framework, as well as the transfer of liability. In the case of a cloud provider, sovereignty and non-reproducibility risk complete the list. Lastly, AI outsourcing amplifies the "blackbox effect", thus making explaining its behaviour more challenging.

Respondents also echo the taxonomy of **attacks against ML** presented in the APCR's document, while also minimising their plausibility of occurrence in a production environment typical for the financial sector. Generic security flaws are deemed more worrisome, such as system intrusion or knowledge acquisition about a fraud detection algorithm (which can then be used for circumvention purposes).

Multi-pronged approach to evaluation

Many respondents adopted their own **analytical evaluation** methods, which focus around the comprehensive documentation of algorithms, of their input data, and of the AI development process, while also addressing the 4 evaluation principles of the ACPR's document. An ideal evaluation process should enable replaying the model and measuring its performance, especially if it is subjected to independent review or external audit.

Data benchmarking is deemed appropriate for internal audit but raises several objections in the context of an external audit (potential biases, reliance on synthetic data, and impact of neglecting the semantics of business data).

Positions are more divergent on **challenger models**: their results are often viewed as uninformative, they are heavy consumers of the auditor's human and material resources as well as of the audited entity's IT resources, and they are – as noted in the discussion document – limited by a lack of standardisation. A method based on coherence analysis is presented as an interesting alternative: its implementation is less complex and its results easier to leverage than challenger models.

The **explanatory methods** most commonly used by respondents are pre-modelling techniques along with some of the post-modelling techniques mentioned in the ACPR's document. Several state-of-the-art methods stemming from very recent research work have been cited, most of which are at an experimental stage and used for non-critical processes (model construction or customer knowledge management) but very few in production. Respondents are sceptical of counterfactual explanations promoted by the discussion document, whose implementation can be challenging due e.g. to non-disclosure practices for computational methods and to their lower acceptance by the customer.

Respondents appear to favour contextualised, tangible explanations which are in line with common-sense reasoning whilst also complying with regulatory requirements, rather than raw, local explanations aiming at higher technical fidelity to the model.

Regulation

According to respondents, AI regulation should fit in a harmonised European legal framework in order to ensure a level-playing field among actors, potentially relying on a certification of external AI components, while always following a risk-based approach.

A normative approach is deemed unnecessary and sometimes prejudicial, as the existing sectoral regulatory corpus enables overseeing AI-induced developments without requiring more specific regulation. For some respondents however, regulatory clarification would help fill in areas of legal uncertainty. Lastly, the evolution of regulation toward a performance obligation – while remaining technology-agnostic – should benefit AI-driven innovation.