Deep Science Publishing
https://doi.org/10.70593/978-93-49910-40-9_8

DeepScience
Open Access Books

Chapter 8: Artificial intelligence governance in finance: Ethics, bias, security, and regulatory compliance in artificial intelligence systems

8.1 Introduction

Artificial intelligence (AI) systems are emerging as sophisticated upright indisputable arbitrators in finance. From automated trading systems, fraud detection, anti-money laundering, credit risk models to chatbots, increasing expenditures on AI systems are expected, propelled by an acute shortage of data scientists and algorithm developers. The allure of big data and the ability of AI systems to extract complex patterns from data are alluring. Ensemble methods such as random forests are prominently used in credit risk and fraud detection. In automated trading, AI methods using methods such as recurrent neural networks are proliferating, leveraging wealth of market data. Applications of AI systems inherently affect human welfare: accessibility to loans, trust in the stability of payment systems, safety from financial fraud. Outcomes determined by AI systems are being questioned: do algorithms freely pour loans to the applicant, or do they drive the rate of suggested purchase? Problems in AI systems arise due to lack of observability, explainability, documentation, and scrutiny (Asatryan, 2017; Chang et al., 2017; Munoko et al., 2020).

Risks appear in multiple dimensions: ethical, legal, financial, and reputational. Where humans are biased, models learn and systematize ingrained biases and discriminate on grounds not exposed to scrutiny. Mocked by an algorithm, decisions are defenseless against the unearthly arbitrariness of a governing machine component when it comes to explainability. This is criminal, given the abstractness of decisions, unlike when a human denies a loan because of ‘lack of proof of income’, and can substantiate ‘the truth of the matters’ (e.g., risk of faking income and a last resort, fragility to shock). Risks commingle and reinforce each other, often at a multi-layer scale (macro/micro). There

is no uniform definition of ‘fairness’, ‘explainability’, ‘interpretation’, or ‘objective’ across institutions, companies, affluences, classes, and societies across countries. Tensions between faculties of law, regulation, audit, data governance, testing, and codes of conduct are which don’t talk to each other, as are experts on interpretability, testing, maintenance, compliance, and so on. With models trained on biased data, black-box ML methods undeserving of scrutiny, ownership, and risk proliferate. On the other end, with the fragmentation of governments lacking unified assessments, testing periods, and uniformity of evaluations, exams and scrutatoires fail to apprehend ground truth on queered a priori evidences (ZestFinance et al., 2017; Rane et al., 2024).

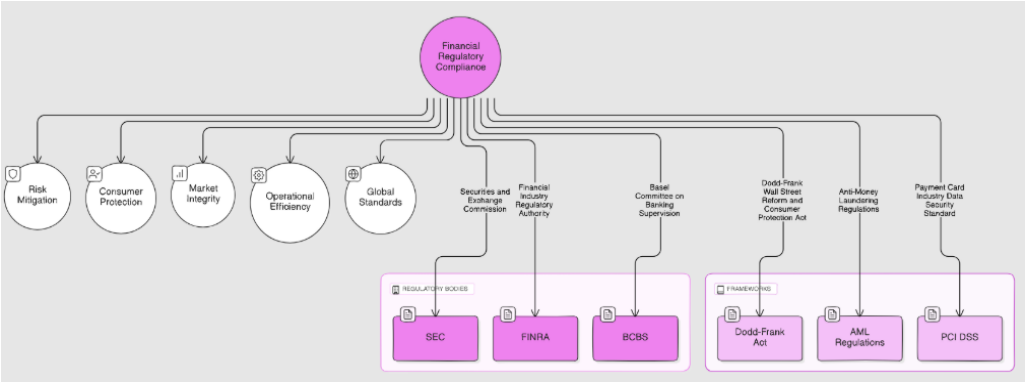

Fig 8.1: AI in Financial Regulatory Compliance

8.1.1. Background and significance

The broad interest in the emerging AI governance landscape has coincided with the rapid rise of AI usage in financial services, particularly in critical systems for which it is vital to preempt and mitigate adverse AI behaviors. Within a reasonable timeframe, financial clients face immense pressure from regulators to demonstrate their AI compliance with regulatory controls and guidelines in a manner similar to examples such as the requirement for EU bank stress tests. A recent survey indicates that over three quarters of global financial institutions have already experienced, or expect to experience, questions from regulators regarding their advanced AI systems within the next two years. Furthermore, it is commonly acknowledged that AI governance is one of the top priorities in many financial services institutions. The speed at which regulatory scrutiny on financial AI usage is being introduced and ramped up suggests that firms need to act quickly to meet the needs for progressive governance. As such, there is a long list of proposed AI governance questions relevant to AI usage in financial services. While these questions and considerations reflect an increasing awareness of the cross-functional nature and multifaceted dimensions of AI governance systems. Any specific technical

solutions to these questions necessarily lie at the intersection of AI systems, application areas, and governance systems.

8.2. Understanding AI in Finance

In finance, AI is often applied in automated trading, credit risk assessment, fraud detection, regulatory recommendation, chatbot, robo-advisory, and others. Core AI techniques include, but are not limited to, supervised learning, unsupervised learning, reinforcement learning, explainable AI, and generative models. Deep learning has been one of the most investigated methods due to its very strong prediction performance and maturity. AI models have been shown to be a state of the art in most of the aforementioned applications. However, AI governance in finance is a largely untouched area. Existing governance systems revolving around regulatory compliance in finance were built with traditional models in mind, and do not easily accommodate systems with drastically different implementations and behaviours. A scarce amount of research has been conducted on AI governance in finance involving financial applications. The sustainable and secure development and deployment of AI systems in finance is situated at the intersection of AI and governance systems.

The long list of AI governance challenges motivates the need for new solution approaches. Researchers present a high-level AI system framework and modular building blocks towards increased self-regulation and more efficient AI governance in financial services. The proposed framework lies at the boundary of the AI and governance systems. The building blocks, once developed, are expected to substantially reduce the implementation cost for AI governance in financial institutions. Continuous regulatory monitoring and reporting during deployment shows the architecture of the proposed AI system framework and regulatory modules. In addition, the research gives a brief technical overview of the enabling technologies for the governance modules. The system aims to incorporate run-time monitoring, regulatory control, and mitigation capabilities in the production environment. Monitoring capabilities are intended for self-regulation purposes. The outputs of the monitoring system tie to regulatory functions to ensure that the system behaves within guidelines.

8.2.1. Overview of AI Technologies

While there is an increasing variety of AI technologies applied in finance, there are also some common AI technologies that appear widely in this domain. The broadest classification occurs along the three dimensions of data type, task, and model structure, followed by a detailed introduction of five common AI model types used in financial

services: Autoencoders, Neural Networks, Support Vector Machines, Decision Trees, and K-nearest Neighbors.

Models may also be broadly classified according to the type of data they work on. Structured data refers to any data that can be stored in a fixed field within a record, having a predefined data model and form. Unstructured data refers to any data that does not have a predefined data model, though unstructured data is harder to analyze than structured data.

Similar to the data types of machine learning systems, the tasks that AI systems are built for may include supervised learning (regression/classification), unsupervised learning (clustering, anomaly detection, dimensionality reduction), and reinforcement learning. The model structure built on these data types and tasks varies. Familiar with modern ML, they can be classified into three classes: tree-based models, kernel models, and deep networks, which will be discussed in detail below. Each common AI model type serves a different type of task, from value estimation to classification, clustering, and anomaly detection. There are also many hybrid models, which are combinations of one or more pure model types, such as deep forest, that combine tree-based and deep network models.

Autoencoders (AE) are unsupervised neural network models, typically including one or more hidden layers. AEs are widely used in anomaly detection by reconstructing data. A pre-trained encoder can also extract useful features from raw natural language text, with only a small labeled data for fine-tuning. In finance, AEs have been applied to identify anomalies in financial transactions.

8.3. Ethical Considerations in AI Governance

The goal of ethics in AI governance should be to incorporate ethical considerations into existing governance frameworks, thereby improving governance processes. This topic encompasses data ethics and AI ethics as they relate to governance, two topics that have been very relevant in public discourse in recent years. Along with governance, risk, and compliance, these concepts were relevant to most financial institutions in the wake of the AI crisis in 2016 and the Cambridge Analytica controversy. AI ethics as applied to governance considers a category of ethical concerns that are relevant to existing governance processes and how these concerns may be addressed. Data ethics as applied to governance concerns which the responsibility for ethical data use should fall within the organization and making a case for better data ethics governance across the board. This approach recognizes that while there have historically been many areas of tension between data ethics principles and governance practice is a category in which the two realms may productively meet.

Some governance frameworks have embraced AI and data ethics by way of codes of conduct that address various concerns. However, coming from entities that are mostly not data- or AI-native, but rather technology followers, these codes have mostly been useless formal exercises, founder-driven crusades to publicly champion positive impact, or "ethics washing" with codes drafted or vetted by PR firms lacking relevant expertise or commitment. Other frameworks have been formed by stakeholder coalitions or consortia that are neither legally binding nor relatively credible. Still others have been developed by organizations analyzing an especially valuable or sensitive data type. However, there is a lack of universal frameworks. Further, even where there are frameworks, organizations frequently lack the requisite capacity to implement or adhere to them and codes may not ultimately matter if governance mechanisms are not credibly enforced against violations. Each of these challenges merits significant systematic examination.

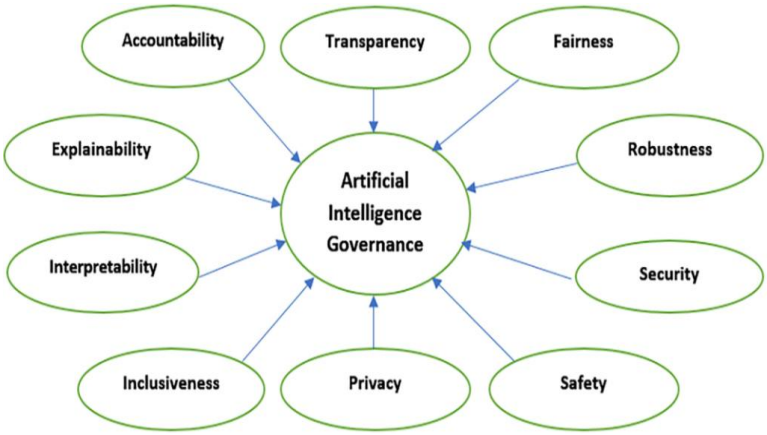

Fig 8.2: Artificial intelligence governance Ethical considerations

8.3.1. Defining Ethics in AI

Many stakeholders have been engaged in defining AI Governance. argue that AI Governance is "the broader activities, policies, structures, and standards that attempt to influence the development, deployment, or regulation of AI systems and technology". This broad definition encompasses a range of institutions, rules, methodologies, and pressures, from formal regulations and laws to informal codes of conduct, norms, and standards. The term AI Governance can refer to anything from formal action against legislation through to advocacy campaigns to lobby for change or to raise awareness, while the very term 'AI' is contested, visioned or understood differently by different stakeholder communities. Many definitions of AI refuse to focus on the specific technologies behind it. A core of analysis of AI Governance is analysing the norms, rules and actions that govern AI technologies. Governance and AI can be understood as

observable actions or behaviours of actors or, in institutional terms, as programmes, processes or structures that govern AI. This sends analysis into actions, events and behaviours of stakeholders undertaking governance of AI. Governance can also have a more social constructionist interpretation.

Law is governance, as is public policy, as are politics and statecraft, but the precise heuristic that decouples the ‘governance’ from the ‘non-governance’ is difficult to operate. To constrain and structure the analysis, it is proposed to focus on AI regulation and individual actors’ actions and efforts of governance that do not take the form of institutions or rules, processes or structures, in short that do not take the formal form of governance (bottom-up governance). Governance actors can be classified as actors representing an indivisible entity or coalition of entities (state actors); actors acting on behalf of an organizational network (intergovernmental actors); actors acting on behalf of a trade or professional association or combining entities across organizations (voluntary actors); or actors representing a personal interest, often acting as individuals on social media platforms (collective actors).

8.3.2. Case Studies of Ethical Dilemmas

A simple example to illustrate the dilemma of balancing AI model transparency against privacy is that many AI techniques offer great power of predicting upcoming events (his/her credit risk) given an individual's prior behavior, while at the same time being extremely challenging to explain exactly how the predictions were made without exposing the underlying data (prior behavior). A very frequent example might be a bank's credit risk prediction model. Almost certainly the prediction would be based on a combination of many variables collectively creating the bank's view of an individual's creditworthiness. However, this credit risk model or its equivalent banks cannot disclose it in detail, lest the banks lose such ability. The immediate concern is societal wealth transfer. A bank winning its fight against impending risks would profit while the losers (eventually defaulted clients) would lose their credits and ultimately their livelihood. An indirect concern was post-fight impact. If the world and human life-style were meaningfully altered by deployment of AI technology, these changes and resulting ethics or morality issues may remain traumatic. Consider again the simple bank credit risk dilemma. Only if the predictions or underlying AI techniques were made public, only with such transparency would regulators and courts have power to judge whether a bank was doing the right thing, or whether the fight was done in a fair manner.

Bank public disclosures of prediction models are generally governed by intellectual property (IP) rights, preventing regulators or attorneys from acquiring sufficient understanding of a bank's prediction model for catalyst AI model technology's ethical scrutiny. Reciprocally, the bank's could self-justify violent actions because such actions

were unavoidable without exposing the model. Therefore, prediction models of AI algorithms should be implemented in a way that strangers or authorities cannot reverse-engineer equation and logic trees, yet predictions can still be independently verified without revealing the model. While costly and cumbersome certainly, it should be pursued, lest AI model developers gain too much political power otherwise.

8.4. Bias in AI Systems

AI systems that make real-life impactful decisions or recommend decisions that get executed by the human user are becoming important components of many applications, including those in finance, healthcare, education, predictive policing, human hiring, etc. In decision-making, such systems should follow various ethical principles, many of which are derived from AI ethics guidelines jointly proposed by ethics scientists and researchers in the AI domain. One of these ethical principles is fairness, which is the absence of bias. The goal of this work is to understand how to recognize bias in AI systems and how to eliminate it when it is found. AI systems that have a substantial effect on people's lives must provide justification and reasoning using appropriate, legally-justifiable, and factually-correct data. There has been much recent interest in assessing the potential impact of AI (in decision-making) on fairness. This concern includes both attacking discrimination against a protected class (e.g., race, gender, religion, etc.) as well as ensuring market competitiveness (such that markets do not devolve into monopoly). The AI systems to be examined in terms of bias fall into one of two mutually exclusive categories: (1) a set of services along with their training data is available, or (2) the AI service is used without access to the trained ML model or underlying training data. In the first case, some bias in the underlying training data or resulting from the training algorithm can be detected, and the designer of the AI service has a covenant to eliminate such bias. Consequently, appropriate debiasing procedures exist. In contrast, if an end-user uses a service without having access to the training data, it must be tested using accessible data that is potentially algorithmically independent of the training data (e.g., historical recidivism data along with current demographic data) to see if the service is biased or discriminatory in any way. There are primarily two distinct approaches to addressing fairness in AI systems. The first approach is to address fairness in AI predictive models before or during the training of models or algorithms, as well as to modify existing models post-hoc.

8.4.1. Types of Bias in AI

Bias in AI systems could be a consequence of pre-trained bias drift or propagation to the downstream model. In case of bias drift, one considers an AI service with some set of

model parameters, such as weights and hyperparameters, trained on a dataset collected from a certain training distribution. This AI service is subject to a new dataset with a different distribution that is obtained under the same schema and procedure. The distributions differ with respect to a certain attribute or group (a subset of the attribute). This leads to the drift in the bias on the AI service with respect to the sensitive attribute. When real-world data is collected and fed to the downstream non-sensitive model ecosystem, pre-trained bias would propagate to the downstream AI services resulting in the bias on the bias-prevented data too. Propagation is not enabled in all AI services. Nonetheless, contemporary AI services in turn build new AI services on top of the existing AI services in a user-friendly manner and thus, inter-service propagation is assured with a high likelihood. The proposed attack is trained with an upper-bound possible bias fully transcribed to provide a bound of the bias.

Ultimately captured an ideal inter-service attack scenario. The rest of the model architecture training does not depend on the client service and thus the non-sensitive training model could be frozen. Parameter freezing seeks to charge up the disinformation-loaded intercept, which could be trained without access to pre-trained parameters. The data are unlabeled and thus trainable input embeddings are appended. The generated embeddings are fed to the held-out client training model for the remaining forward pass. The training procedure adopts the standard optimizer and the loss function as follows. Hyperparameter setting was used to count the weight decay factor while infrastructure specifics are omitted. After 30 epochs, the mean square error with the frozen non-sensitive training model plateaus at approximately 1.3.

8.4.2. Impact of Bias on Financial Decision-Making

First, AI-based systems for financial decision-making present clear, sequentially, and easily defendable reasoning processes. In the case of the financial services industry, such model clear-as-day reasoning is often required by regulators and critically important to firm risk management; industry practice prefers transparent and yet explainable choices. This is one important reason why tree ensemble methods and linear regression are so-overly-abundant in financial AI. Second, some AI-based systems are too complex and esoteric to have any externally understandable reasoning processes, even though this does not mean the model does not produce results that population-level, longitudinal data identifies to be valid. For example, neural networks are often considered a hard-cut black box, and lack of clear reasoning is acknowledged by practitioners. Aside from reasonable concerns raised with AI bias assessments in general, a specific challenge with hybrid AI and ‘subsystem’ bias detection has been identified commonly encountered by finance practitioners within the finance context and out-perform standard benchmark bias metrics. This recommendation strongly emphasizes quantifiable reasoning over

qualitative concept drift. Bias detection and treatment for AI-based systems devoted to financial decision-making, and cases where AI-based systems led to biased financial decision-making. Hybrid AI bias statement describes the (exact) nature of the problem and action taken (AI-based system, model specification, etc.) to treat the concerning bias following recommendation aspect structure.

Explicitly highlight design features in both the statement and socio-technological conditions around the AI system that by design mitigate biases, and other factors that were either unconsidered in design or that are deemed as forces beyond control in the given socio-technical condition. Most AI-based systems devoted to financial decision-making tend to have excessive model input characterisation for precautionary (regulatory) purposes. News-based variables and custom factors designed to detect crowd behaviours are a further reflection of the modelling choices made in view of the external environment and were often constructed to reduce financial market impact risks instead of personal bias consideration.

8.5. Security Concerns in AI Applications

AI technologies have the potential to aggrandize challenges regarding telecommunication security. Therefore, it is necessary for financial institutions to assess the security of their AI systems, either received from third-party vendors or in-house developed after deployments. Financial institutions should ensure that data is protected appropriately throughout the AI system life cycles, including data obtained, retained, used, and destroyed. In addition, financial institutions should carefully consider their use of public data. To ensure a thorough understanding of the risks involved, financial institutions should solicit independent third-party expertise if they choose to recreate public data themselves. As more sophisticated AI technologies emerge, data may persist even though it appears not to exist, creating risks that may not be relevant for conventional techniques. To preempt data breaches regarding data held by service providers, financial institutions should conduct appropriate vetting of potential service providers, emphasizing security, handling, and deletion. AI systems should limit data retention to what is necessary for the AI model. For AI models that require retention beyond the typically accepted limits, reasonable vindication practices should be constructed when determining whether data used for algorithmic decision-making can be accessed.

AI systems should be additionally supervised and safeguarded against potential risks, training or applicational, to avoid delayed remediation. Financial institutions may obtain independent third-party technical experts to conduct comprehensive reviews of AI models. Risks are unlikely to be effectively remediated unless the initial state-of-the-art model can be recreated. To ensure this ability, the pre-requisites specified in the above

seventh guideline should be observed to granularly explain the model, guarantee reproducibility if requested, and assign responsibility for its creation. Furthermore, financial institutions should implement internal monitoring and safeguards to ensure the coherence between the deferred outcomes/decisions and the standards of appropriate care. Financial institutions should add monitoring and remediation capabilities, including remedies for input tampering, exploration in production, and robustness checking. Fundamental research that yields insights into this class of technologies is crucial to grounding and refining regulatory safeguards and in doing so may also recommend other concepts that deserve inclusion in a comprehensive framework of governance.

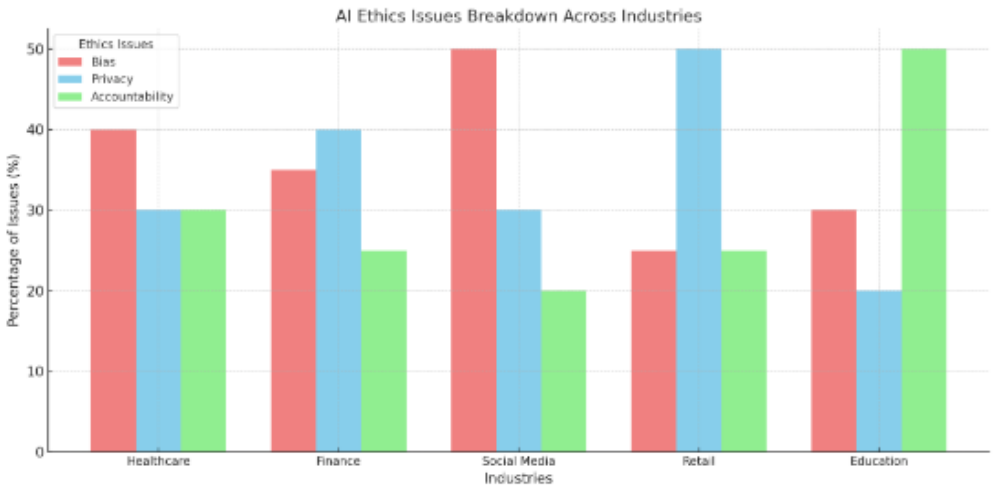

Fig : AI Ethics in 2025

8.5.1. Vulnerabilities in AI Systems

AI systems have undoubtedly been of great interest to scholars as they offer new opportunities to improve and automate business practices. However, the rapid adoption of novel deep learning approaches in financial services has led to some concerns. Most prominent questions include robustness, compliance with pre-existing regulations and laws, ethical, unintended bias, algorithmic discretion, transparency, and interpretability. As AI systems become an increasingly integrated part of the financial services industry, missing an effective model governance workstream may result in a great number of unexpected consequences such as bias, over-reliance on technology, model failure, and potential regulatory repercussions. AI warfare adversarial attacks are even more aggressive and harmful than those in a non-AI context since they can happen in a higher frequency, larger scale, and more unexpected forms.

These unexpected consequences of AI systems can originate from a number of factors. Most fundamentally, models can behave in an unexpected manner not only because of model weaknesses, but also due to underlying data and assumption weaknesses. AI algorithms such as adversarial learning also have a greater degree of combinatorial explosion not present in traditional methods. Meaningfully explaining the effects, working mechanisms, and rationale of models is more difficult given their complexity. Further, some existing model governance practices in the financial services industry may not be applicable to some AI models given their structure, assumptions, and detrimental consequences. With the unprecedented growth in AI model complexity and the difficulty in black-box models, the feasibility of existing model governance practices for the next generation of AI enabled, explainably, and interpretable model is in doubt. One way in addressing this feasibility issue is to make the model governance process more automatic and interpretable so that human experts can efficiently assess them without an overwhelming workload. The standard of the assessments can be made more flexible so that input from experts working in different fields can be gathered. AI-enabled tools can also be utilized under human-on-the-loop guidance to increase model and governance capabilities.

8.5.2. Data Privacy Issues

The accumulation of analyses and information gained from implicit mechanisms raise distinct consequences on privacy issues [5]. Privacy concerns arise when financial AI systems require the activity, value, and personal data of clients, clients' companies, and their counterparts. After gathering the relevant data to form the models for AI processes, third parties may obtain sensitive data on companies' risk assessment and creditworthiness as well as the exposure of clients' portfolios to various market and credit events. This knowledge can be leveraged by various actors to generate vulnerability or create abrupt market failures in the financial services industry. Additionally, the financial activity and data of consumers may also find a way to be revealed. The disclosure of such information can result in serious implications on any acts involving financial activity because they would be, in effect, subject to the scrutiny of others. Synthetic data can be provided in the modelling processes but there exists a significant potential bias based on the construction and statistical properties of such data. Thus, any form based on non-aggregated financial data should be at the period preceding that of financial crises. Nevertheless, there is no guarantee of the absence of any bias prevalent in an environment approved by authorities.

8.6. Conclusion

AI, machine learning and other technologies capable of accelerating the analysis and understanding of large volumes of structured and unstructured data and their interdependencies, and predicting their movements, behaviours and valuations, are already extensively applied in finance. Financial firms increasingly rely on AI systems for various decision-making processes, some of which may affect large groups of people and communities in important ways. ML systems are prevalent, for example in credit and insurance risk assessment, bias monitoring to improve assessment fairness, anti-money laundering compliance detection, customer service through chatbots, equity trading strategy development and backtesting, and algorithmic enhancement of more traditional statistical risk models. While the capabilities and benefits of AI systems are promising, especially for some finance domains that are hierarchical and non-linear in nature, they are often considered associated with more risk and blinder than traditional algorithms.

Unlike traditional algorithms, many AI systems by design operate as black boxes. This is particularly true for state-of-the-art, deep learning systems, which are composed of many heavily interacting layers that adjust their parameters over millions of observations. AI systems are often opaque to their own users. The rigidity and lack of interpretability associated with these systems lead to challenges for risk understanding, model validation, regulatory compliance, and even simply linking model input and output. In finance, the stakes are high when things go wrong or when a model stops working well, and the degree of error permissible is far lower than in applications such as speech recognition or smart advertising. At the same time, because ML is a rapidly evolving technology, the institutional and regulatory landscape is far behind in understanding these models and the risks and trade-offs associated with their application. Regulators have not provided guidance on risk-relevant features and tests, and explanations required for interpretability. Litigators struggle to challenge or defend algorithmic decisions made within a black box.

8.6.1. Future Trends

The long list of AI governance challenges motivates the need for new solution approaches. Existing governance efforts are desired to be more capable, adaptable, and efficient. This section envisions and presents a high-level AI system framework and modular building blocks towards increased self-regulation and more efficient AI governance in the financial services sector. The proposed AI system framework aims to support self-regulating capabilities via run-time monitoring and enforcement of compliance, while the governance building blocks, encapsulating common regulatory requirements, are designed to be integrated as modular units in AI systems towards

efficient compliance by design. To present and understand the proposed architectural framework, operating AI systems in financial services are first examined. AI systems are centralized machine learning deductive models that are broadly instrumented to be integrated in a Decision Supporting System. A rigorous machine learning development process, also commonly implemented, composes a modeling environment, where AI models are built, validated, and peer-reviewed before deployment, and a production environment, where production model management and lifecycle monitoring are performed. The federated AI system framework architecture brings together a comprehensive perspective of how solutions are executed with its components in connection, the inputs and outputs of its components, the alerts employed for feedback control, and the governance considerations at each step of the process. Systematic misbehavior scenarios are also summarized based on their relevance to financial services. Essentially event- and regulation-condition triggered checks make the operation of the governance packages and regulatory monitors highly flexible and efficient and ensure only requisite events are captured and addressed. These checks can be asset-level, group-level, and time-intensity conditioned, while the severity levels can also differ. Focusing on regulatory package design and integration considerations, the proposed framework is expected to better prepare and regulate AI systems in compliance with customization and frequent updates of the packages. The regulatory consideration is also relevant to pre-regulatory design compliance in managed development environments.

References

Rane, N. L., Choudhary, S. P., & Rane, J. (2024). Artificial Intelligence-driven corporate finance: enhancing efficiency and decision-making through machine learning, natural language processing, and robotic process automation in corporate governance and sustainability. Studies in Economics and Business Relations, 5(2), 1–22. https://doi.org/10.48185/sebr.v5i2.1050 sabapub.com+1ResearchGate+1

Chang, H., Kao, Y.-C., Mashruwala, R., & Sorensen, S. M. (2017). Technical inefficiency, allocative inefficiency, and audit pricing. Journal of Accounting, Auditing & Finance. Wikipedia

Munoko, I., Brown-Liburd, H. L., & Vasarhelyi, M. (2020). The ethical implications of using artificial intelligence in auditing. Journal of Business Ethics. Wikipedia

Asatryan, D. (2017). Machine learning is the future of underwriting, but startups won't be driving it. ZestFinance. Wikipedia

ZestFinance. (2017). Zest Automated Machine Learning (ZAML) platform for credit underwriting. ZestFinance.