CAPCO

MANAGING AI RISKS
IN FINANCIAL SERVICES

a **wipro** company

The themes of human agency, fairness, accountability, auditability, and transparency are central to the subject of autonomous decisions made by AI, and the recent high-profile emergence of ChatGPT has brought the topic of AI ethics firmly into the spotlight.

Financial services organisations (FSOs) are currently unprepared to manage their operational exposures to this emerging type of risk. With legislation becoming inevitable, this paper discusses key considerations to ensure AI can be effectively governed, and AI-related risks identified and mitigated, via a proportionate control framework.

Ultimately, the goal of impending legislation will be to protect fundamental rights and enhance trust in autonomous decisions made by AI. FSOs stand to gain significant early mover advantages by proactively meeting the challenges.

Over the past year, it is has been difficult to maintain even a passing interest in technology without reading about and experiencing the remarkable capabilities of ChatGPT and other generative AI applications available to the public. The invention of generative algorithms has pushed AI technology to the forefront of public consciousness, with its potential to significantly transform our society and economy. However, this potential comes with well documented risks that governments will not be able to ignore. The risks specific to financial services organisations can be broadly split into two categories:

1. **External risks** arising from usage of AI by an external party to subvert internal controls such as those within the Anti Financial Crime units of banks. The ability of AI to pass as 'human-like' will only progress with time, exposing FSOs to increased risk of financial crime, insurance fraud, and cyber security breaches, for instance through phishing attacks and identity theft.
2. **Internal risks** arising from an organisations own use of AI due to potential breaches in ethical codes and associated regulatory compliance. The maturation of non-deterministic AI technologies in applications such as chatbots and credit scoring raises stakes not only in the ability of AI to do more than just handle repetitive tasks and routine queries, but also in the risks of doing so. As autonomous decisions made by AI become more widely adopted, there are increasing ethical ramifications around human agency, oversight, fairness and transparency.

How organisations should address the challenges posed by internal threats stemming from their own use of AI is the primary focus of this paper.

In the EU, the Ethics Guidelines on Trustworthy AI[1] published in 2019, set out a framework of principles and requirements for ensuring AI is ethical, transparent, trusted and has legitimate basis. Compliance with these guidelines is voluntary and uptake has been slow to non-existent in financial services.

However, while many current applications of AI are benign, this fast maturing technology could come to pose threats to our society, economy, the environment and even our personal safety. This makes legislation inevitable in the near to medium term that will impact all sectors of the economy. At the time of writing, the draft Artificial Intelligence Act in the EU has been overwhelmingly approved by the EU legislative branch and is now moving through its final stages towards enactment[2]. Even if FSOs have been slow to respond historically, legislation will compel the industry to act.

In this context, those that invest proactively in their data and control environments will gain huge comparative advantages. Early movers will nuture an AI-centric culture that can both leverage AI and control the associated risks at greater economies of scale, compared to those that scramble (at great cost) to do so only when the law requires it.

WHERE ARE AUTHORITIES FOCUSING?

Ethical concerns associated with increasing innovation in the use of technology and data is already driving governmental bodies to require FSOs embed data ethics into their risk management frameworks. The best example is the EU's General Data Protection Regulation (GDPR) which enshrines the ethics of data privacy into law. Another example is the UK Consumer Duty Act, within which data ethics is a key theme, that requires FSOs to evidence the delivery of fair outcomes for all.

The draft EU AI Act that is currently being reviewed by member states appears on the surface to target public services and law enforcement, but there are some holistic takeaways for FSOs in the risk-based approach taken:

Critical / unacceptable risk categories that will face an outright ban, covering usage of AI in:

- Social / behavioral scoring – **e.g. credit scoring**
- Subliminal manipulation of behaviour – **e.g. through marketing of financial products**
- Exploitation of vulnerable population categories – **e.g. differential pricing of risk**
- Emotion recognition – **e.g. in hiring or fraud investigations**

High-risk categories requiring review, certification and post-market monitoring, covering usage of AI in:

- Biometric identification – **e.g. usage of biometrics in insurance pricing**
- Employment and human resource management
- Predictive policing of anti-financial crime
- Assistance in legal interpretation and application of the law – **e.g. in privacy decisioning**
- Educational access – **e.g. in determining access to training**

Limited risk categories requiring transparency measures covering:

- Spam filters
- Chatbots

It should be noted that generative AI models such as ChatGPT may fall into any of the above categories, depending on the context and nature of application.

While the EU approach is likely to involve the enactment of AI principles in centralised, AI-specific legislation, the approach in the UK is currently devolved, sector-specific and guidelines-based, leveraging existing legislation in specific industries and in parallel legislative domains such as data protection. Within the UK financial services, the Bank of England (BoE) and the UK Financial Conduct Authority (FCA) have led efforts to provide additional guidance to FSOs[3,4], focusing on the following impact areas:

- **Consumer protection** – e.g. usage of AI in provisioning of credit
- **Fairness of competition** – e.g. price collusion as an unintended outcome of AI
- **Insurance policyholder protection** – e.g. usage of AI in biometric identification for purposes of tailored insurance
- **Financial stability** – e.g. usage of AI automated trading
- **Market integrity** – e.g. usage of AI in fraud detection and susceptibility of FSOs to AI-enabled fraud
- **Operational soundness** and safety in data management, model risk management and AI governance.

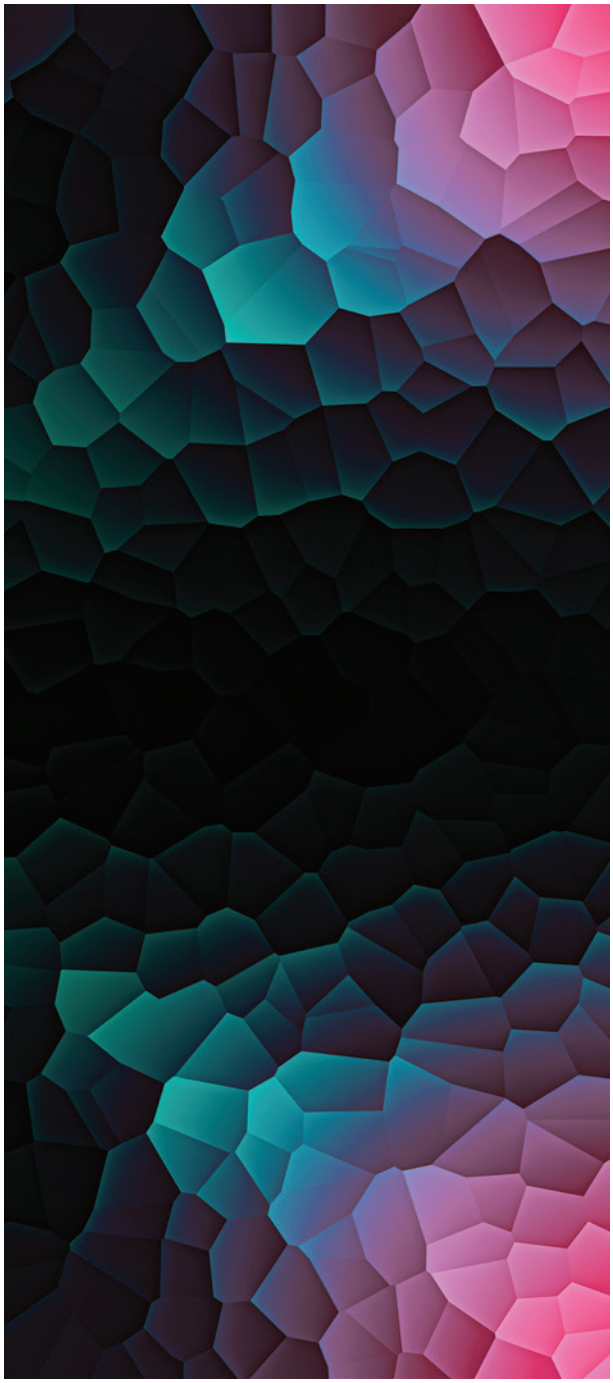
Generally, it is important to recognise that AI-related risks are regarded as being fundamentally ethical in nature. As such, the regional approach to legislation is likely to be similar in nature to the treatment of other ethical issues such as data privacy, as shown by the contrasting approaches adopted so far in the EU and the US.

In the US, despite the introduction of the Algorithmic Accountability Act[13] in 2022 the approach to AI legislation is likely to be via state legislature[11] although this is not to say the Federal Trade Commission (FTC) may not move to legislate in response to the EU taking a leadership position in global AI regulation. However, in the near term at least, there is more likely to be a patchwork of rules across the US, much akin to the current approach to privacy – it should be noted that several states including California already partially legislate for AI-related risks through existing data privacy frameworks.

In summary, the global regulatory landscape is as varied and evolving as quickly as the technology itself, with the emergence of generative AI representing a step change in technological capability towards the emergence of Artificial General Intelligence (AGI) agents that can actually think for themselves. The exponential development and advancement in both technology and associated risks puts governing bodies under increasing pressure to regulate.

In this context, leaders at the 2023 G7 summit strongly signalled the desire for more international collaboration and alignment on AI governance. The following legislative possibilities (non-exhaustive) should therefore be considered and FSOs, especially those with a multi-national footprint, should prepare accordingly:

- **Restriction or outright ban** on certain types or use-cases of AI in certain jurisdictions or industries
- Requirement to **certify and register** AI applications with one or more regulatory authorities
- Requirement for riskiest AI applications to be **externally or independently monitored**
- Requirement for **transparency** on the role and extent of AI involvement in interactions and decisions made
- Requirement to subject AI outputs to **internal or external review** (e.g. developed in a regulated sandbox) post-market
- Requirement to demonstrate **security and robustness** in the design, development and operation of AI models
- Requirement to protect the organisation from **copyright infringement**
- Requirement for **failsafe mechanisms** to ensure processes can be easily be stopped or switched over to fully manual processes should red flags be detected



WHAT DOES THIS MEAN FOR FINANCIAL SERVICES ORGANISATIONS?

The competition imperative should be driving a scramble by FSOs to build their AI capabilities – a blanket ban on AI is not a competitive option either for individual firms, or for economies as a whole[5]. However the risks are very real and will only become increasingly more so. FSOs should therefore ensure global AI strategies are established that not only aim to scale and industrialise the ability to leverage AI (such as establishing curated data platforms), but also take proactive steps to mitigate current and future AI-related risks.

This will include:

- Establishing strong **data management** foundations to ensure quality, availability and trust in the data inputs
- Maintaining a **global inventory**, classification and tagging mechanism for AI models
- **Controlling for bias** in the operational lifecycle of AI (e.g. MLOps), provisioning for:
  - Maintaining model stability
  - Minimising the risk of unfair bias in outputs
- Improving the **transparency** of outputs
- Embedding an **algorithmic impact assessment** (AIA) in the control environment, incorporating the EU's proposed conformity assessment:
  - Assessing algorithmic risks and determining required mitigants both a-priori and throughout the AI lifecycle
  - Imposing location-based restrictions on access to and usage of specific classifications of AI
  - Ensuring AIA decisioning is both *risk and value* based

FIGURE 1:
AI RISK MANAGEMENT FRAMEWORK

- Implementing **governance and oversight** to internally police the development and usage of AI, and manage, monitor and report related risks. This will cover:
  - Establishing policies and standards on AI development, usage and control
  - Articulating and embedding AI-related risks into the operational risk framework
  - Establishing appropriate governance and regulatory response bodies
  - Defining, designing and implementing metrics and Key Risk Indicators (KRIs)
  - Establishing and empowering human oversight of AI
  - Monitoring of model drift and unintended outcomes
  - Defining training and certification pathways for AI to be trusted with higher risk functions.

In the following sections, we highlight some of these key areas for further scrutiny.

ESTABLISHING STRONG DATA MANAGEMENT FOUNDATIONS

At Capco we have consistently called out the importance of establishing strong data management foundations in order to build effective data science capabilities[6]. AI is no different in this regard – badly managed data not only raises the marginal cost of AI, but also causes proliferation of unamanaged data quality issues that drive statistical bias in AI outputs. Conversely, the ability to curate well-defined, governed, trusted, quality data in a cost effective manner enables scale economies in the production of fair and transparent AI, driving the potential to leverage AI at industrial scales.

MAINTAINING A GLOBAL INVENTORY

Management of AI-related risk will require an up-to-date and complete knowledge of where these risks reside. Furthermore, if different types of AI (e.g. neural network vs decision trees) carry different risks, FSOs need to identify and distinguish their AI models within an agreed taxonomy/ontology to ensure policies can be consistently applied at granularity. These considerations require FSOs to maintain a complete inventory of their AI models covering:

- Taxonomic/ontological classification
- Controlling/owning legal entity and/or jurisdiction
- Business function
- Scope of decision-making e.g. routing logic, product recommendation, pricing
- Scale and scope of usage (geographic and functional)
- Data input classifications
- Risk rating.

MITIGATING THE RISK OF UNFAIR BIAS

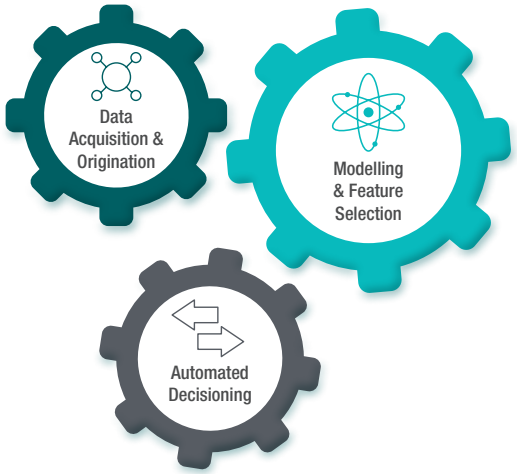
Successful, law-respecting businesses do not intentionally set out to infringe upon human rights – ethics risks rather lie in the unintended consequences of analytics/AI usage. Some of the key considerations around mitigation of these risks involve:

- **Mitigating unfair bias introduced via the training data** – AI engineers need to ensure curated training data sets are representative of the population and free of embedded human subjectivity, bias, or errors. As these biases are typically subconscious, independent expert review is typically required as part of the AIA (Algorithmic Impact Assessment) and with each training iteration.

- **Mitigating unfair bias in specifics of feature set, model construction/parameters, and from model drift** – During model validation, unfair bias is checked for both in the feature set chosen and with the impact of model parameters. Features and parameters can be further refined, or additional rules/blacklists introduced, to ensure outputs meet ethics policies. Model validation is a key activity both on initial go-live and in post production monitoring to correct model 'drift'. In ML, the relationship between target and input variables change ('drift') over time, unlike traditional analytics where this relationship is static and specifically point in time. ML models require refreshing over time which drives a need to continuously monitor, maintain and validate the model, even after initial go-live.

- **Mitigating unfair biases that arise from non-deterministic behaviour of AI algorithms** – Model validation by itself is necessary but insufficient to mitigate the risk of unintended and unethical outcomes arising from non-deterministic algorithms. This requires the output to be independently monitored (proportionally to the assessed risk) and reviewed. The governance bodies (internal or regulatory) may require that all output from high risk AI be reviewed before that output can be used in production.

FIGURE 2:
KEY SOURCES OF UNFAIR BIAS



IMPROVING TRUST AND TRANSPARENCY

Transparency around decisions involving AI models is key to maintaining **trust in AI**, whether with the public or regulators. Given the increasing complexity and non-determinism of AI models, the risk that decision-making transparency involving AI models will be eroded only increases.

Transparency means it must be possible to technically trace AI decisions back to the data that was used to train that decision. Techniques such as the maintenance of replication files and design/change logs with each AI iteration should be adopted to maintain clear **traceability** from input data through, the decisioning processes, to the model outputs.

Although traceability is a necessary condition for transparency, it is however not sufficient by itself. Part of the transparency objective is that the outcomes and how they are derived can be **explained** to business stakeholders and regulators. This requires a high degree of ability to bridge technical understanding of the technical trace with the business context in which AI decisions are made. **XAI (explainable AI)** is achieved through the shift from 'black box' to 'glass box' decision making[7] in this manner.

Explainability is still a significant technical challenge however, especially for more opaque techniques that show non-deterministic, emergent outcomes such as deep learning/neural networks. There are three key considerations in this regards:

- **Not everything is explainable, or needs to be explainable, to everyone – experience counts**. Medical decisions may not be explainable to the general public in detail, but nevertheless, we are more likely to trust a senior doctor than a junior one. Similarly, trust in any specific AI instance can be built over time, evidenced through certification and a history of production usage. AI should be internally certified on simpler/less risky goals to begin with, and then gradually applied to increasingly complex and risky functions over time with additional certification.

- **The mimum standard is explainability to subject-matter experts**. Building on the medical analogy, medical decisions should be explainable to other medical professionals. Although the behaviour of many models is too complex to describe in simplified business terminology to the non-expert, the process of documenting and explaining the model in an inventory means that it is auditable and open to scrutiny by other subject matter experts.

- **Collaborative, multi-disciplinary research**. FSOs should dedicate more resource to researching XAI methods as it is key to unlocking AI use cases, giving significant competitive advantages to those that can lead and innovate in this area.

RISK AND VALUE BASED ALGORITHMIC IMPACT ASSESSMENT (AIA)

The objective of an AIA is to determine the mitigating actions required based on the **overall** level of ethical risk a-priori. It should take place at the initiation of all AI projects and can be implemented as a natural extension of a Data Protection Impact Assessment (DPIA).

The EU's Ethics Guidelines for Trustworthy AI offers a full articulation of some of the questions that might be posed in an AIA. In summary, the broad considerations are:

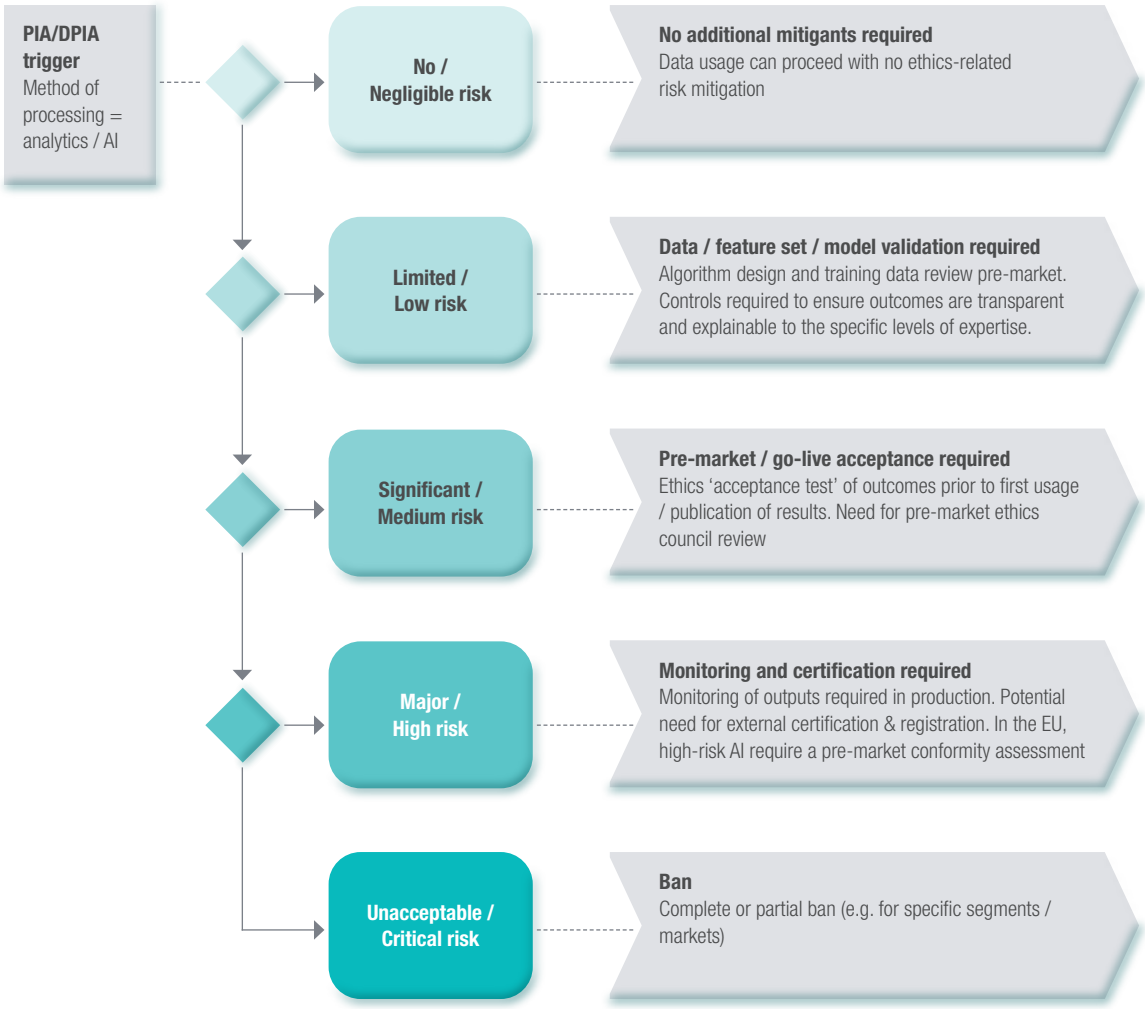
- The extent to which humans will retain agency and autonomy in decision making
- The impact on society and welfare – e.g. will jobs be at risk?
- The impact on fundamental human rights – e.g. impact on privacy
- The extent to which the AI will interact with humans, and the nature of interactions
- Measures taken to improve and validate technical robustness and security
- The reliability, reproducibility and transparency of outcomes.

FIGURE 3:
OUTLINE FLOW AND OUTCOMES OF AN ALGORITHMIC IMPACT ASSESSMENT (AIA)

Rules-based risk assessment → **Required mitigation**

Expanding upon the PIA/DPIA, the AIA assesses ethical risks in relation to the intended usage of data. Questions are structured to create the shortest path to assessment outcomes, enabling lower risk projects to exit the assessment earlier.

Not all mitigations are required for all projects. The AIA instructs projects of the required controls that are commensurate with the level of risk identified, with the least risky projects requiring no additional control and the riskiest projects potentially subject to an outright or partial ban.



Although the AIA should be configured so that a 'high-risk' outcome is obtained for any AI system categorised as 'high-risk' by the EU, the AIA should be sensitive to applicable legislation in other jurisdictions.

Once a level of risk is determined, appropriate mitigating measures can be recommended:

- The least risky projects may not need to do anything more than an initial risk assessment
- Moderately risky projects may additionally require a pre-production review of data, feature set and model bias
- The riskiest projects, short of being banned, may also require external certification and monitoring in production

Note that it is also important to consider the value of use-case being proposed in order to better inform a balanced decision – at Capco we often find that organsiations expose themselves unnecessarily to AI/ML risk through use-cases that do not generate a commensurate amount of value. Therefore a pro-innovation approach in line with UK government principles[8], would additionally involve the balancing of risk with value, prior to any decisioning.

GOVERNANCE & OVERSIGHT

Mitigating ethical risk requires ethical judgements to be made, and the need for human oversight is central to governing AI well.

In this sense, AI governance is much more than just the management of model metadata and ensuring outcomes are reproducible. It must also encompass the controls and processes to ensure AI has **human oversight**, that governance structures are in place to track, manage, escalate and report AI-related ethical risks, and that AI is managed in production in a manner proportional to risk.

Human oversight – As decisions become increasingly delegated to AI, decision makers will need to adjust habits and perceptions to accommodate augmented and automated decision making. Critically, where decisions become fully automated, humans will move from a position of responsibility for decision making to one of accountability. That is, **decision makers will become decision owners**. This has several implications:

- The functional and operational scope of delegated decision making needs to be clearly articulated and tracked from requirements through to production of AI models
- Decision owners bearing accountability for augmented and automated decisions of AI models, including any adverse outcomes thereof, will need identifying as part of an overarching governance structure for AI
- Decision owners will need to become actively involved in the training and testing of AI models
- Decision owners will need to be proactive in monitoring AI outcomes – the delegation of decision making to an AI does not release humans from responsibility of the impact of those decisions
- Decision owners will need to establish a deeper understanding of the AI-driven decision-making process, and the risks and potential causes of adverse outcomes in order to fully discharge their function.

The last point especially, indicates the need for AI strategies to be accompanied by a **cultural shift** involving proportionate training and skill uplift not only for technical staff, but also for those **on whose behalf AI will be working.**

Governance structure – While introducing some automation in the decision-making process, especially at lower levels of risk, reduces operational overhead, human accountability for AI outcomes is an imperative. This should be reflected in the AI governance structure, and appropriate policies and processes should be put in place to ensure active participation in AI governance by those responsible and accountable for decision making.

However, it is not necessary for a such a governance structure to be monolithic and centralised. Indeed, doing so risks over-burdening the organisation, adding unnecessary costs, and reducing the pace of innovation. The organizational structure of governance should be tailored to suit the level of risk, with only the highest levels of risk requiring governance at the highest levels of accountability.

At the highest levels of governance, an operating model for interacting with external regulatory bodies will need to be considered to register, certify and monitor high-risk AI developed by the organisation.

Post-market AI management – Organisations will need operating models to manage AI in production, particularly around:

- Management and monitoring of risk throughout the AI lifecycle to protect from model drift and the risks associated with unintended outcomes
- Ensuring AI-augmented business processes can be made to 'fail-over' to manual processes should AI instances start exhibiting unintended behaviour
- Training 'pathways' for AI to be internally certified with perfoming higher risk functions
- Regulatory horizon scanning and management
- Incident reporting and management of breaches in data ethics and adverse AI outcomes.

CONCLUSION

Without a doubt, generative AI is a game changer, and some would argue is as revolutionary as the internet[9].

However, this potential is not without serious risks to fundamental human rights, social cohesion, safety, law and order, and national security. The likelihood of regulation is therefore very high. However there are many aspects of AI governance and risk management for which most FSOs, even those that have efficient and effective ethics mechanisms, are currently unprepared.

The emergence of ChatGPT and other generative AI models is a reminder to FSOs that the rapid pace of AI innovation compels firms to continuously review and update the relevant risk and control frameworks accordingly. They can put the building blocks of such a framework in place by:

- Strengthening their data management operations
- Inventorising their AI models
- Mobilising an operational lifecycle for AI that mimises the risk of unfair biases and improves transparency
- Embedding an Algorithmic Impact Asssessment to ensure AI model risks are understood, a priori
- Establishing effective governance structures and human oversight over AI.

The impact of poor risk management and controls in this space are manifold but could include missed sales opportunities, litigation, regulatory fines, reputational damage, and social detriment, violating corporate social responsibilities. Even if governments and regulators do not step in until AI risks are manifesting at severity and scale, individual FSOs should not leave becoming the trigger for such regulation to fate. Quite the opposite – they should be proactively investing in both leveraging AI techniques and in controlling the related risks to build long-lasting comparative advantages in AI usage.

REFERENCES

1. European Commission, “Ethics Guidelines for Trustworthy AI,” April 2019. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

2. European Parliament, “MEPs ready to negotiate first ever rules for AI,” 06 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai.

3. Bank of England, “AI Private-Public Forum Final Report,” February 2022. [Online]. Available: https://www.bankofengland.co.uk/-/media/boe/files/fintech/ai-public-private-forum-final-report.pdf?la=en&hash=F432B83794DDF3F580AC5A454F7DFF433D091AA5.

4. Bank of England, “DP5/22 – Artificial Intelligence and Machine Learning,” December 2022. [Online]. Available: https://www.bankofengland.co.uk/prudential-regulation/publication/2022/october/artificial-intelligence.

5. NikkeiAsia, “Banning Generative AI is not an option for Germany: minister,” April 2023. [Online]. Available: https://asia.nikkei.com/Spotlight/G-7-in-Japan/Banning-generative-AI-is-not-an-option-for-Germany-minister2.

6. Capco, “Data management - A foundation for effective data science,” November 2019. [Online]. Available: https://www.capco.com/en/Capco-Institute/Journal-50-Data-Analytics/Data-Management-A-Foundation-For-Effective-Data-Science.

7. AI4Good.org, Lindsey Asis, “What is explainable AI and how does it apply to data ethics,” May 2021. [Online]. Available: https://ai4good.org/blog/what-is-explainable-ai-and-how-does-it-apply-to-data-ethics/.

8. UK Government, “Establishing a pro innovation approach to regulating AI,” July 2022. [Online]. Available: https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement.

9. Business Insider, “Bill Gates calls ChatGPT ‘every bit as important as the PC’ or the internet,” February 2023. [Online]. Available: https://www.businessinsider.com/bill-gates-chatgpt-ai-artificial-intelligenct-as-important-pc-internet-2023-2?r=US&IR=T.

10. European Commission, “Artificial Intelligence Act (draft),” April 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN.

11. Alston & Bird, “AI Regulation in the U.S.: What’s Coming, and What Companies Need to Do in 2023,” December 2022. [Online]. Available: https://www.alston.com/en/insights/publications/2022/12/ai-regulation-in-the-us#:~:text=U.S.%20AI%20Regulation%20to%20Expect%20in%202023,-The%20rapid%20growth&text=Several%20use%2Dcase%2Dspecific%20AI,and%20new%20NIST%20AI%20standards.

12. Scientific American, “AI drug discovery systems might be repurposed to make chemical weapons,” April 2022. [Online]. Available: https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/.

13. United States Congress, “Algorithmic Accountability Act of 2022,” Jun 2022. [Online]. Available: https://www.congress.gov/bill/117th-congress/senate-bill/3572/text.

AUTHOR

Alvin Tan, Principal Consultant

CONTRIBUTOR

Chris Robinson, Principal Consultant

CONTACTS

Alvin Tan, Principal Consultant, alvin.tan@capco.com
Matthew O'Kane, Partner, matthew.okane@capco.com

ABOUT CAPCO

Capco, a Wipro company, is a global technology and management consultancy focused in the financial services industry. Capco operates at the intersection of business and technology by combining innovative thinking with unrivalled industry knowledge to fast-track digital initiatives for banking and payments, capital markets, wealth and asset management, insurance, and the energy sector. Capco's cutting-edge ingenuity is brought to life through its award-winning Be Yourself At Work culture and diverse talent.

To learn more, visit www.capco.com or follow us on Facebook, YouTube, LinkedIn and Instagram.

WORLDWIDE OFFICES

APAC
Bangalore – Electronic City
Bangalore – Sarjapur Road
Bangkok
Chennai
Dubai
Gurgaon
Hong Kong
Hyderabad
Kuala Lumpur
Mumbai
Pune
Singapore

EUROPE
Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Milan
Munich
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA
Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Washington, DC

SOUTH AMERICA
Alphaville
São Paulo

WWW.CAPCO.COM

CAPCO
a wipro company