European Risk Management Council

Risk Landscape Review

December 2023

- Towards a new model risk management paradigm
- Risk Sentiment Index: Update for the UK and the US

DEAR READER,

I am delighted to present the Q4 2023 edition of the Risk Landscape Review, which includes two articles.

The first article is dedicated to model risk management in the era of AI expansion. In this article, William W Hahn, Partner and Global Head of Financial Risk and Model Risk Management at TCS CRO Strategy Advisory, analyses challenges of model risk management for complex models like Generative AI and Large Language Model. The author concludes that to effectively manage the model risk of AI-driven models, a new model risk management and governance paradigm should be adopted.

The second article is dedicated to Q4 2023 update of the Risk Sentiment Index (RSI) for the UK and the US, an expert-driven, forward-looking index that reflects the expectations of experts about the risk landscape of the financial sector in the next 12 months. The results of surveys recently conducted in the UK and the US suggest that while Chief Risk Officers and other risk decision-makers in both countries expect a relatively mild increase in risk in 2024, the potential areas of stress and volatility might differ for the UK and the US financial services.

My huge thanks to all contributors and survey respondents. Enjoy the reading.

Yours sincerely,

Dr Evgueni Ivantsov

Chairman of European Risk Management Council

Table of Contents

Generative AI (GenAI) and Large Language Models (LLM)- Towards a New Model Risk Management Paradigm

By William W Hahn, CFA, Partner and Global Head, Financial Risk and Model Risk Management, TCS CRO Strategy Advisory

Managing AI and ML Risks—Transparency and Data

Although regulators in the U.S. (Federal Reserve and OCC) have not promulgated any regulations governing AI/ML directly, they have in the past stated that AI/ML can be used by so long as the banks have in place the process for identifying and managing potential risks associated with AI or ML as with any models used in the bank. Over the years, MRM validations and regulatory examination of AI/ML "models" have tended to focus on two primary sources of risks: lack of transparency and data accuracy and representativeness.

AI and ML are black box algorithms. To provide some intuition or "story" behind how these models work, MRM community developed techniques and methodologies collectively called "Explainable AI" or XAI. There are two approaches to XAI, and both attempt to explain or interpret indirectly how ML is using the inputs to produce the output. One approach attempts to look at all the data inputs (features) and determine how significant or important the features were in determining the output (prediction or classification). The other approach attempts to reverse engineer a more interpretable model around a local feature space. An example of this approach involves sampling data near a prediction and fitting a local linear model that is subject to interpretation. These and similar methodologies are ways to build trust and buy-in from stakeholders and to intervene to change the outcome.

Data representativeness and accuracy also represents big source of risk because of the heavy dependence of the AI/ML algo on data. The quality of the outputs depends on the sufficiency and the quality of the data used to train the models. If the data itself is inaccurate or not representative or if there is a pattern that suggests bias or discrimination, the ML algorithms will simply pick up on those signals or patterns and perpetuate it. Using data sample of sufficient quantity and quality to train and validate the AI algorithm ensures better performance, and statistical tools and techniques exist to mitigate bias in the algorithm. Such techniques may include over or underweighting certain data observations to address skewed data or other imbalances (for e.g., too much or too little credit decisions made for certain classes in the data sample). Other bias mitigation techniques include ensuring that measurable or observable targets being optimizing (i.e. surrogate objectives) are truly indicative of legitimate business objectives (for e.g. minimizing hospital visit costs may not truly predict or assess health risks of individuals and may discriminate against low income minorities), and changing the parameters or optimization routine to ensure a "fair" outcome, (i.e. that a protected class has no better or worse outcome than other classes all else being equal). These and other similar methodologies have been described extensively and have been used widely in addressing data accuracy and bias risks.

What Does "Interpretable AI" or "Bias Mitigation" Mean for GenAIs and LLM?

Large Language Models (LLMs) and other Generative AIs (GenAIs) represent the next evolution in AI/ML technology. GenAIs are considered general-purpose AI designed not for any single task or function, but as an interface to allow greater accessibility to complex and technical subject matter and to accomplish tasks quickly. LLM's power and applicability come from its seeming ability to "comprehend" human language and to provide a coherent response to queries concerning any topic. Yet, questions remain whether the tools and methodology the MRM community has developed to manage the transparency and data risks are adequate or even relevant for managing risks inherent in GenAIs.

Like all AI/ML, GenAI's and LLMs are giant black boxes but with complexity that far exceeds anything that has been used in banking and finance thus far. No one, not even the programmers at OpenAI themselves, knows how ChatGPT has configured itself to produce texts that is so human like—its neural architecture is not based on any real theory or engineering. So, transparency is a real issue with these exceptionally large and complex GenAIs. Unfortunately, no interpretable AI/ML has ever been tried on a neural network the size and complexity of an LLM the likes of ChatGPT. And even if interpretable model were theoretically and computationally possibly, one must ask whether concept likes "linear approximation" or "feature importance" are even applicable to an AI that generates content. What are the important features when the only features are numerical vectors representing words or word fragments? What we call "interpretable" or "explainable" AI currently is meaningless in the context of GenAIs.

The most pertinent use case for a general-purpose AI like ChatGPT is gathering and summarizing information. Yet, despite its human like writing ability, ChatGPT is a statistical machine. It is providing "plausible" or "probable" responses based on what it has seen during its training. It cannot verify or give reasons for its responses. The random nature of the LLMs also makes them susceptible to "hallucinations"—flights of fancy that LLMs pass off as being true especially when broached with a topic for which it has received little training data. So, accuracy and veracity are significant issues with GenAIs. But what does "bias detection and mitigation" mean when all LLMs are doing is generating content? Are we talking about content generation that is tilted towards certain language or ethnic groups (English versus say, Swahili) or producing illegal, unethical content (dissemination of hate messages or publication of copyrighted materials)? The concept of "bias mitigation" as it is used currently is meaningless when applied in the context of GenAIs.

A New Model Risk Management Paradigm

Given the lack of progress on "interpretable" AI and a clear definition of "bias mitigation" in the context of GenAI's, there is a compelling case to be made for a whole new set of model risk management policy, one that acknowledges that GenAI's are self-organizing algorithms whose process for generating content cannot be decoded as a mathematical or physical model (at least not yet) and one that focuses not on "how" they work or "what" they are doing but "what" they are

generating and "how" they are used in the organization. In this paradigm, the model risk governance is geared towards verifying whether the GenAI is being used within the organization in a safe and responsible way, and whether there are adequate internal control checks in place to ensure the veracity and appropriateness of the AI content when the business lines rely on it as part of their daily activities. It is a paradigm where the second and the third lines of defense are not so much validating the AI model but auditing the business user's justification and rationale for relying on the LLM's outputs and attesting to the adequacy of the safeguards and protocols in place to ensure legal, ethical, and responsible uses.

Methodologies for Safe and Responsible GenAIs

Safety and responsibility are not just about governance; they can also be made an intrinsic part of the GenAI programming themselves. LLM's parameters can be better tuned and refined using additional curated domain specific set of training data from both the third-party data vendors as well as the firm's private database. This will lead to more accurate, relevant, and institution-specific responses to user requests and queries. Writing better, more detailed queries (prompt engineering) or providing contextual definition within each prompt (for e.g., a predefined set of functional uses cases from within which users can only write their prompts) can also improve the accuracy and relevance of the LLM's responses. LLMs can also be forced to provide cites to the sources from which it draws its responses so that the users can verify and cross check the information contained in the responses.

Apart from using better data and engineering prompts, training techniques may be layered in to provide further assurance of ethical and responsible uses. "Reinforcement learning from human feedback" (RLHF) is one such technique. In RLHF, LLM responses are graded by humans based on the appropriateness of the responses. The human feedbacks are then fed back into the neural network as part of its training. The goal is to reduce the likelihood of the LLM providing a harmful or outright untrue statements when given similar prompts. Another technique is "red teaming." In it, users simulate "attacks" by writing bad prompts to get LLM to do what it should not be doing in anticipation of malicious activities that can happen out in the real world. The goal is to identify those sets of "bad" prompts and to set guardrails around the responses. The last technique involves using another AI to police the LLM responses. In this technique, a secondary neural network, which has been trained on ethical and legal principles or policy, is used to monitor the responses for compliance with legal and ethical principles. This secondary AI can also be used as part of the initial LLM training itself. The goal is to have the LLM optimize for the most ethical and legal response when considering its choice of the most relevant and accurate responses.

Conclusion

GenAIs represent a huge leap in AI/ML technology that have democratized accessibility to complex and technical subjects and analytics. Model risk methodologies developed and applied in the model risk management of AI/ML have become irrelevant in the age of super complex AI's and mass adoption of AI technology. A new model risk management and governance paradigm should be adopted—one that establishes use standards and enforces safe, ethical, and responsible uses.

Risk Sentiment Index: Q4 2023 Update

Views from both sides of the pond

In Q4 2023, the European Risk Management Council launched the US Risk Sentiment Index (RSI), complementing two similar indices already produced by the Council for the UK and APAC. It allows for the comparison of risk sentiment in the financial services sector across several jurisdictions. Chief Risk Officers and other senior risk executives from banks and other financial institutions provided their views on future trends for seven types of risk, including credit, market, liquidity, operational, cyber & IT, conduct, and regulatory risks.

RSIs represent numerical interpretations of the adjusted percentage of respondents who anticipate an increase in risk over the next 12 months. Consequently, a higher RSI indicates that more respondents expect an upswing in risk.

The fresh data collected for Q4 2023 from the UK and US provide insights into the perceptions of UK and US executives responsible for risk management functions regarding dynamic trends within the risk landscape.

UK Risk Sentiment Index: Key findings

The aggregated RSI for the seven risk types has risen from 0.42 in Q1 2023 to 0.49 in Q4 2023, signalling a shift towards a more pessimistic sentiment among the surveyed respondents (Figure 1).

The shift towards a more pessimistic outlook is reflected in the distribution of votes among the five voting options (Figure 2). The proportion of respondents expecting a significant increase in risk over the next 12 months has climbed from 19% in Q3 2023 to 29% in Q4 2023. Conversely, a percentage of respondents who expect no change in risk declined from 31% to 26%. Overall, these changes suggest that respondents foresee heightened volatility in the UK financial services sector throughout 2024.

Analysing the quarter-to-quarter changes in the RSIs for individual risk types reveals an increase in six out of the seven risks (Figure 3). Only RSIs for market risk experienced some reduction compared to the previous quarter. Notably, liquidity risk recorded the most significant quarter-to-quarter increase.

Despite some fluctuations quarter-to-quarter, cyber risk and credit risk remain the focal points of concern, with respondents anticipating significant challenges in managing these risks over the next 12 months. The RSI for cyber risk currently stands at 0.74, the highest among all seven risk types, followed by the RSI for credit risk at 0.62 (Figure 3). Respondents harbour worries about the potential escalation of credit risk driven by the challenging macroeconomic conditions and high interest rates in the UK. But in spite of an increase in Q4 2023, the RSI for credit risk stands within the long-term average (Figure 4).

Figure 1. UK RSI trend: Q4 2018 – Q4 2023

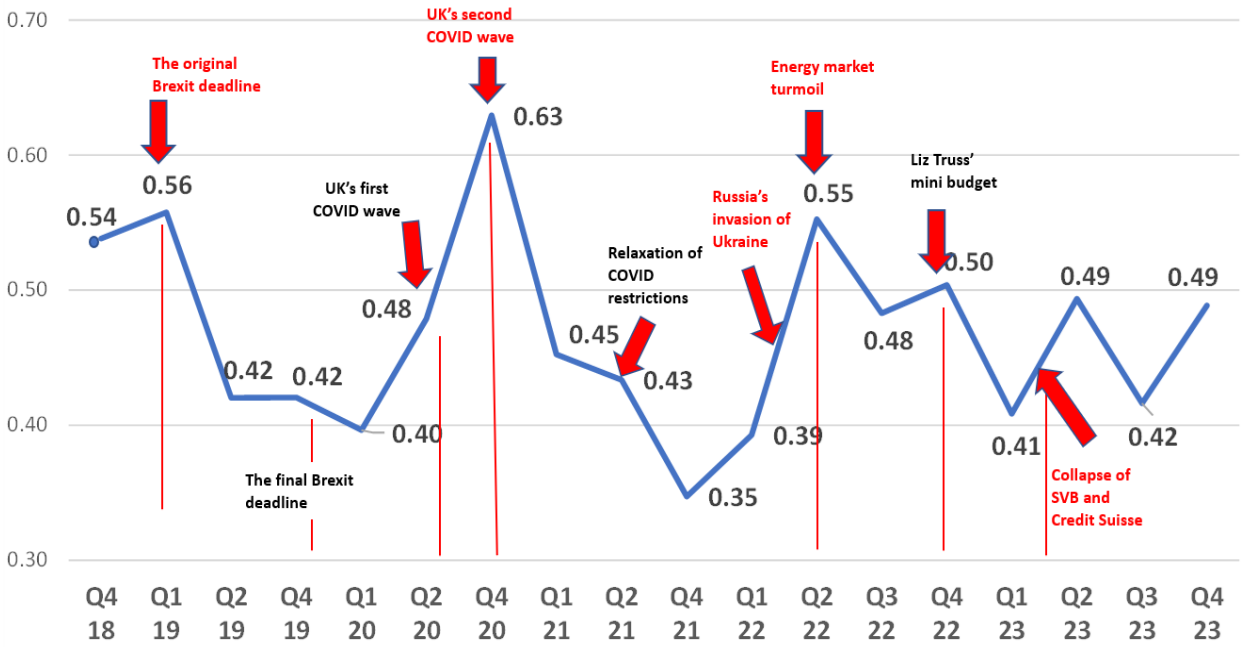


Figure 2. Distribution of respondents' votes

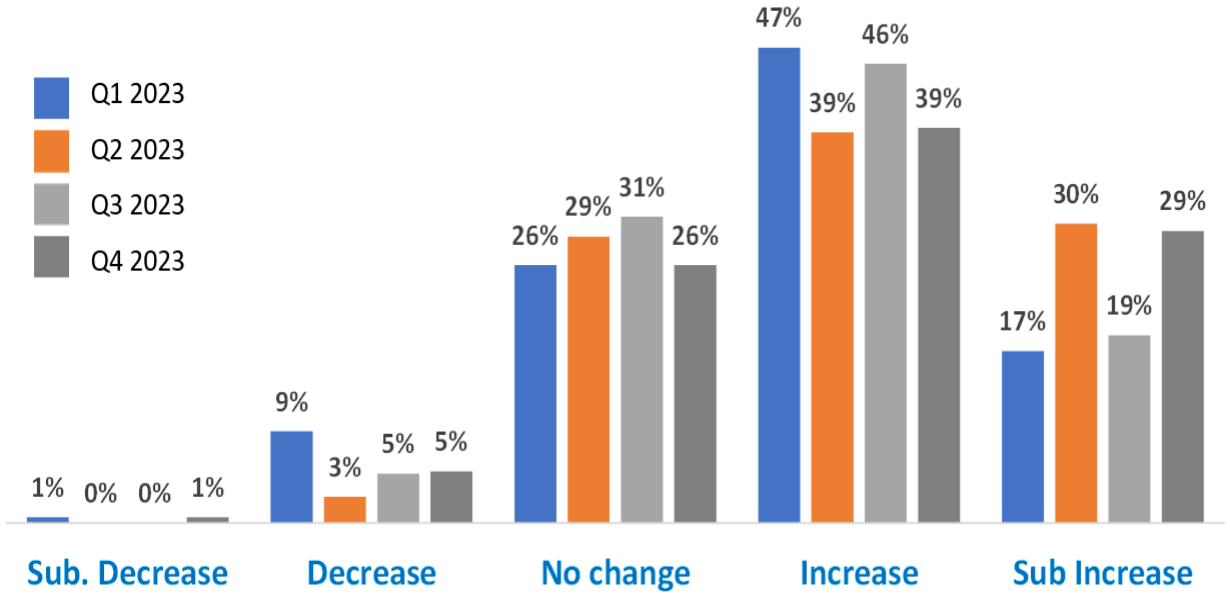


Figure 3. Recent RSI trends for individual risk types

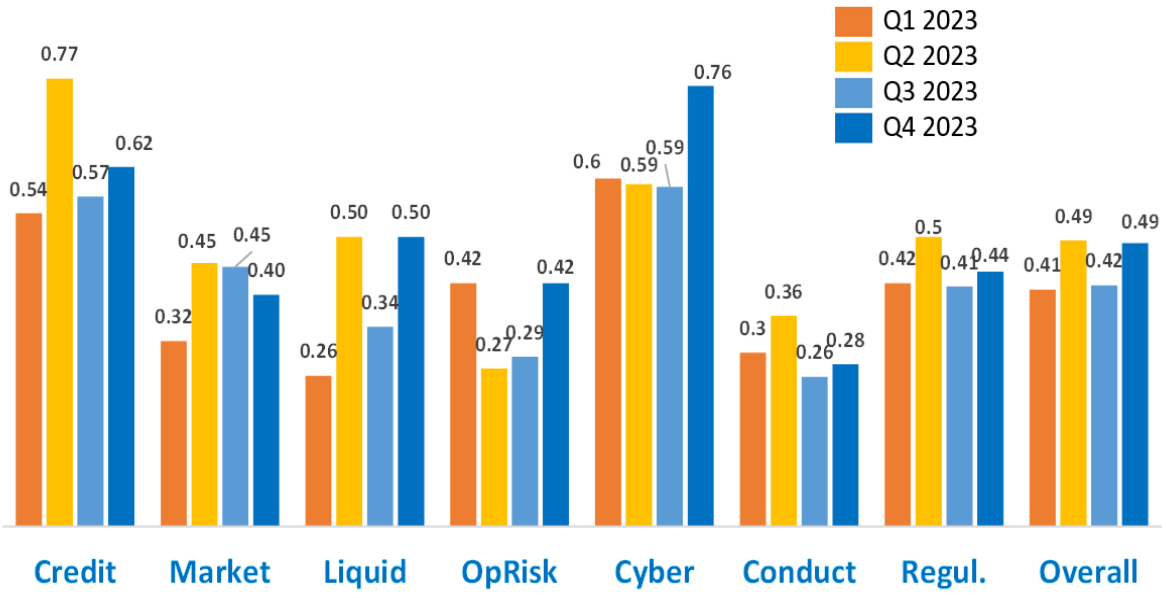
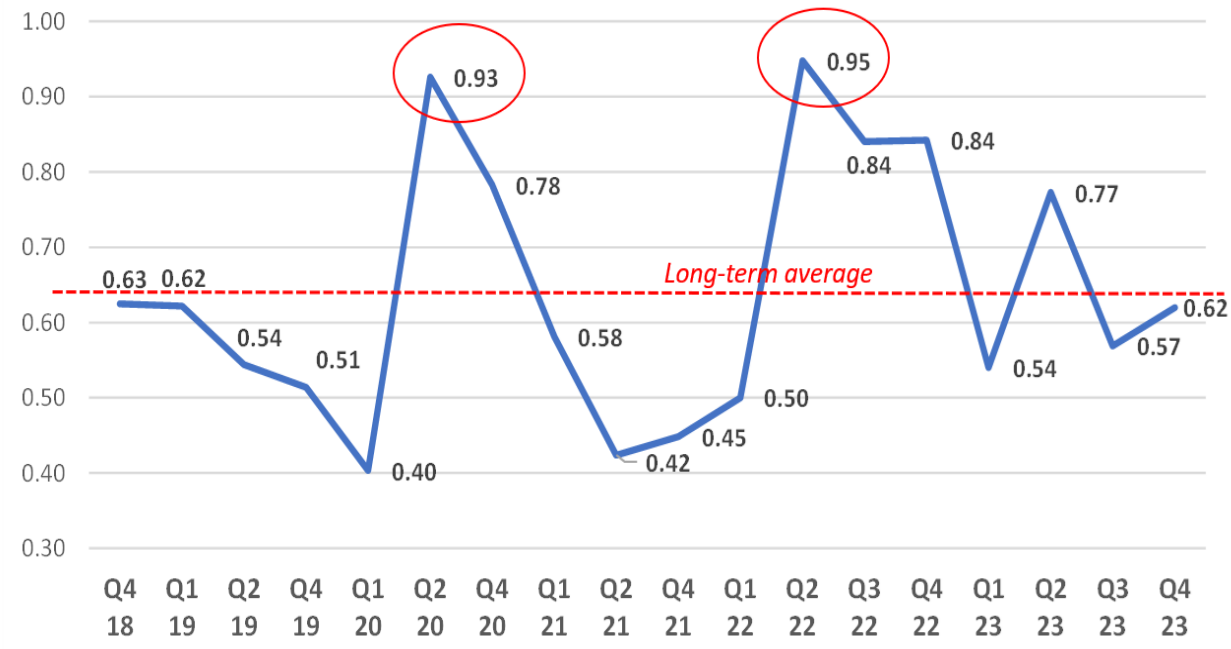


Figure 4. UK RSI: Credit Risk Trend Q4 2018 – Q4 2023



US Risk Sentiment Index

The aggregated US RSI for seven risk types is 0.46 in Q4 2023. This figure is quite close to the aggregated UK RSI of 0.49. A comparison of the distribution of respondent votes across five voting categories reveals a close similarity, with 68% and 69% of respondents expecting risks to increase in the UK and US surveys, respectively. Additionally, roughly a quarter of respondents in both countries anticipate no change in risk over the next 12 months (refer to Figure 5). Overall, these numbers indicate that respondents in the US and the UK expect a relatively mild increase in risk.

US respondents consider credit risk to be their primary concern, anticipating a more significant increase in credit risk compared to other risk categories in the next 12 months. The primary reason for this concern is the high interest rates on the US dollar, and if they persist for an extended period, they may undermine the financial resilience of debtors, leading to higher credit losses next year. On the other hand, UK CROs are more concerned than their US counterparts about the potential growth of cyber risk (refer to Figure 6).

Another substantial difference between these two national indices is CROs' sentiment regarding regulatory risk, highlighting the distinctive regulatory environments in these countries and the varying approaches of the two nations' regulators. For US CROs, regulatory RSI is the second highest after credit RSI, suggesting that respondents anticipate substantial pressure from US regulators.

The US regulatory RSI stands at 0.57, with 80% of respondents expecting an increase in regulatory risk in the next 12 months, and a third of respondents believe this increase will be substantial. This concern arises from the tightening of regulations by the Fed and FDIC in the aftermath of the 2023 mini banking crisis. Specifically, a new, more stringent regulatory capital regime known as Basel III Endgame is expected to be implemented next year.

In contrast, regulatory risk is not a top concern for UK respondents. The UK regulatory RSI is ranked number 4 only among the 7 risk types. The regulatory RSI for the UK currently stands at 0.44, roughly in line with its 5-year average.

Figure 5. UK vs US: Distribution of respondents' votes

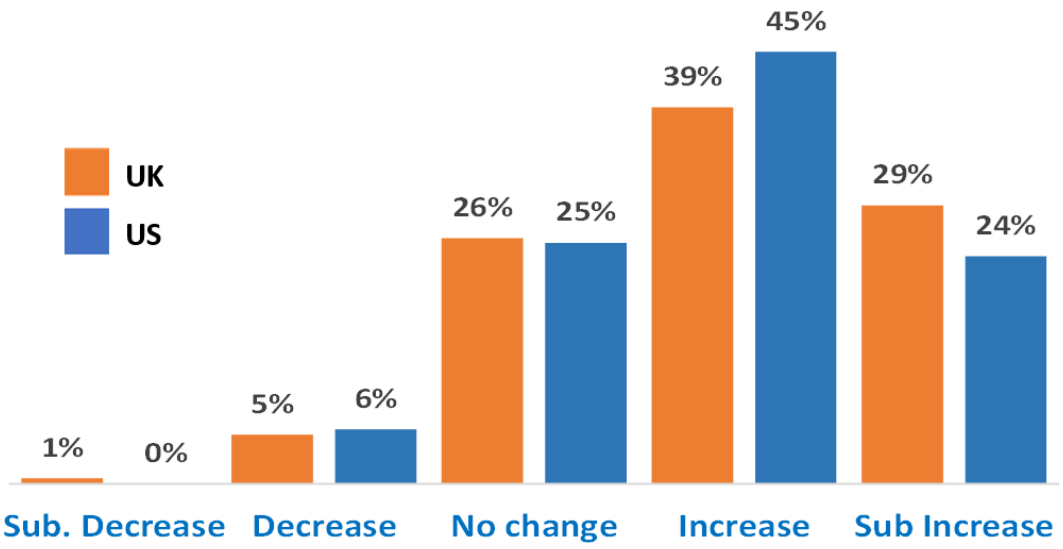


Figure 6. UK vs US: Comparison of RSIs for different risk types (Q4 2023)