Request for Information: Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning

June 28th 2021

- Docket ID OCC-2020-0049

  Chief Counsel's Office
  Attention: Comment Processing
  Office of the Comptroller of the Currency
  400 7th Street SW
  Suite 3E-218
  Washington, DC 20219

- Docket No. OP-1743

  Ann E. Misback
  Secretary
  Board of Governors of the Federal Reserve System
  20th Street and Constitution Avenue NW
  Washington, DC 20551

- RIN 3064-ZA24

  James P. Sheesley
  Assistant Executive Secretary
  Attention: Comments-RIN 3064-ZA24
  Federal Deposit Insurance Corporation
  550 17th Street NW
  Washington, DC 20429

- Docket No. CFPB-2021-0004

  Comment Intake
  Bureau of Consumer Financial Protection
  1700 G Street NW
  Washington, DC 20552

- Docket No. NCUA-2021-0023

  Melane Conyers-Ausbrooks
  Secretary of the Board
  National Credit Union Administration
  1775 Duke Street
  Alexandria, VA. 22314-3428

Subject: Request for Information - Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning

Dear Administrators,

We welcome the opportunity to comment on 'Financial Institutions' Use of Artificial Intelligence, including Machine Learning'. We organized our response as follows:

A. Introduction to Regulatory Questions

B. Kernel Surrogate Functions

C. Regulatory Questions

I. Explainability (Questions 1 to 3)

II. Risks from Broader or More Intensive Data Processing and Usage (Questions 4 and 5)
III. Overfitting (Question 6)
IV. Cybersecurity Risk (Question 7)
V. Dynamic Updating (Question 8)
VI. AI Use by Community Institutions (Question 9)
VII. Oversight of Third Parties (Question 10)
VIII. Fair Lending (Questions 11 to 15)
IX. Additional Considerations (Questions 16 and 17)

D. Concluding Comments

E. Authors (Dr. Robert Mark & Dr. Gary Nan Tie**)**

Please share our response with your respective regulatory agencies.

A. Introduction to Regulatory Questions

Our response to the Regulatory Questions is guided by the notion that an AI model needs to be clearly explained and fit for purpose.
AI models are used to explain data as well as to make predictions. Data is usually collected from observations or experiments. Associated with each data point is usually a label value. In other words, each input is associated with an output.

One challenge in making predictions from an AI model is to construct a function from inputs to outputs which behaves reasonably based on existing inputs and will make plausible predictions of outputs on as yet unseen inputs. The reasonableness and plausibility is where parsimonious model selection comes in.

A parsimonious AI model choice embodies an economy of conceptualization wherein we avoid either being unnecessarily elaborate or being too simplistic. We choose a parsimonious AI model in terms of only that which is needed to understand our problem and robustly extrapolate. By doing so, we better understand and mitigate model risk.

Parsimony is context dependent. What is complicated in one framework may be simpler in another. For example, duality is the notion of a bidirectional relationship that upon round trip is somehow equivalent to the starting point. Sometimes a dual problem is easier to solve than the original. A prototype example is the use of Laplace transforms in solving differential equations. So, while everything should be as simple as possible and not simpler it should also be context dependent

The right definitions are crucial in creating context. To paraphrase Manin[1], having the right definitions is more important than having proofs because results are almost obvious with the right conceptual framework. Parsimony can sometimes be achieved by having the right perspective. What is considered right evolves over time just as consensus on a correct mathematical proof has evolved over time.

We hope that our response provides value for all interested stakeholders who need to address AI model risk. We recognize that various stakeholders have different analytical capabilities and therefore have structured our response to reach out to all stakeholders independent of their analytical skills.
We introduce the idea of surrogate analysis and are guided by this theme since it naturally threads together in a coherent way our responses to each of the 17 Regulatory Questions.

We prepared a Q&A to further elaborate why surrogates provide a common ground as follows:

Q1: What is a surrogate?

[1] Manin, Y., (1998),' Interrelations between Mathematics and Physics', Societe Mathematique de France.Manin states that "All the other vehicles of mathematical rigor are secondary [to definitions], even that of rigorous proof."

A1: A surrogate is a function that parsimoniously approximates an objective function of interest, like a black-box AI algorithm.

Q2: What does a surrogate do?

A2: A surrogate is designed to make sense of what an objective function is doing, especially if it is complex, computationally expensive, or proprietary.

Q3: Why use a surrogate?

A3: By understanding what an objective function is doing, a surrogate further enables examination of an objective function's predictiveness, robustness and fairness.

Q4: What are surrogates used for?

A4: Surrogates are used for:

- vetting black-box model performance, predictiveness, robustness, and fairness
- optimizing hyperparameters
- counterfactual extrapolation
- stress testing
- monitoring model drift

- managing model risk
- model comparison and benchmarking

Q5: Why not use a surrogate all the time?

A5: The discovery of data patterns and their fine detail is what an objective function achieves. Surrogates complement this by enabling the vetting and understanding of objective function results.

The strength of linkage varies considerably between an explanation and what an AI algorithm actually does. In a perfect world we would have a scientific theory with testable hypotheses for our problem being calculated by an AI algorithm. Unfortunately, this is not always the case and therefore we make do with approaches that we now critique.
A weak form of explanation is an analogy that seeks to compare partial significant similarities. Slightly stronger are heuristics which are simple rules of thumb learnt by experience. Both analogies and heuristics are cognitive processes without a testable connection to an AI algorithm, like for example statistical inference.

Statistical inference is good for helping us understand general tendencies but assumes we have some knowledge of underlying

probability distributions. Another caveat is that correlation is not necessarily causation. Moreover, knowing average behavior does not guarantee that an AI algorithm on a specific instance will act that way.

By contrast, surrogates of AI algorithms have a strong mathematical connection with provable properties. This transparency enables AI vetting by consumers, vendors and regulators, through common understanding of what an algorithm is doing. Surrogates also allow us to pose specific 'What if? questions', stress test AI algorithms and assess the fairness or bias of AI results. Since a surrogate is a mathematical model of an AI algorithm there is model risk, that is a discrepancy between surrogate and AI output that needs to be understood. Parsimonious model selection that is fit for purpose is crucial.

Analogies, heuristics, statistical tests, and mathematical surrogates are all forms of explanation of AI black box results. We need to be aware of their differences in linkage strength, as we try to understand AI algorithms.

B. Kernel Surrogate Functions

AI learning algorithms typically construct the best fitting function from inputs to outputs from a given class of functions[2]. So, although we can explicitly articulate the steps of an algorithm, we may not necessarily have intuition about the resultant function chosen.

Recall familiar linear regression, fitting a line through data points, with dependent variable y, independent variables x_i and coefficients a_i, i = 1,…, n:

$y = a_1x_1 + a_2x_2 + \ldots + a_nx_n$

A nonlinear generalization of similar form is a function given by a surrogate formula as follows:

$f(z) = a_1K(x_1, z) + a_2K(x_2, z) + \ldots + a_nK(x_n, z)$

where z is an element of the domain of the black box algorithm and kernel function K(x, z) is a domain specific measure of similarity between data points x and z. Note that f is a weighted sum of nonlinear functions with known well studied properties, which we call a (kernel) *surrogate.*

We have generalized the linear function x_i to a special nonlinear function $K(x_i, \mathbf{.})$ that has properties enabling f to be a surrogate for any black-box algorithm F, that is to say, $f(x_i)$ is designed to approximate $F(x_i)$ well, at

[2]For example, deep learning neural networks

all the data points x_i. Simply put, surrogate function f behaves as if it were the black-box algorithm F on the data sample. Even though a black-box algorithm F may be complex or proprietary it's surrogate f however is a transparent explainable proxy that can readily be computed. Most importantly, by understanding what the surrogate does, we can make sense of what the black-box algorithm is doing.

Specifically, given just a sample of AI algorithm input-output pairs x_i, $y_i = F(x_i)$, we first calculate its surrogate f to understand what the algorithm F is doing, which is designed so that $f(x_i)$ approximates y_i well. Then to make sense of the algorithm F, we examine the relative positive and negative contributions of the weights and special nonlinear functions in the formula for the surrogate f. Domain knowledge experts can then interpret and attribute significance to the larger contributions. This is how we can understand and explain AI algorithms.

In the context of credit lending, proprietary black-box algorithms are sometimes used to make credit decisions. Regulators have used logistic regressions in the form of logit functions to evaluate such black-box

decisions[3]. The use of a logistic regression here is an example of using a proxy. Our surrogates given by the formula above is more general.

Our surrogate functions are a parsimonious choice of a data model on a spectrum of model complexity that we can explain. Our surrogates fall between simple models like linear regression, and complex models like deep learning networks. Surrogates help us understand what black-box algorithms are doing. This enables assessment of fairness and bias, as well algorithm comparison and benchmarking. Surrogates provide common ground for users, vendors and regulators to systematically vet AI algorithms and mitigate model risk.

One qualification about the usefulness of surrogate functions beyond their faster computation is that they are most needed in black-box situations where we have little knowledge about how the results were generated. If on the other hand we had a scientific theory about how the data is generated then a surrogate function may not add much insight. For example, suppose we had an AI options pricing algorithm, then beyond being less expensive to compute, a surrogate function

[3] For example, default probabilities are sometimes modeled as a logit function for ease of explainablity as opposed to trying to explain more sophisticated and better approaches to calculate default risk.

may not be needed to understand and explain what the AI algorithm is doing because there is already a well-developed theory of options pricing.

C. **Regulatory Questions**

I. Explainability (Questions 1 to 3)

AI algorithms solve optimization problems that can be articulated but often leave users without a sense of why and how. XAI, so called 'Explainable AI', endeavors to explain AI results in intuitive ways but by doing so adds another layer of risk to modeling as heuristics are not what the model actually does and could be misleading. A more parsimonious and practical way to understand AI would be to do sensitivity and stress tests of algorithm results. These tests involve careful perturbations of input data and examination of how the resulting output is affected

LUSI (Learning Using Statistical Invariants), Vapnik's 2018 foundational learning framework, can be used to perturb predicates in order to reduce VC (Vapnick Chervonenkis) dimension[4]. For example, one can

[4] Abstract of a talk by Vapnik, V., Izmaloilov, R.,(2018),' Rethinking statistical learning theory: learning using statistical invariants', available at https://www.csail.mit.edu/event/learning-using-statistical-invariants-revision-machine-learning-problem

try different sets of predicates to see how VC dimension is reduced. The advantage here is that this form of cause and effect can be intelligently refined.

Given an AI algorithm's output, how can we understand and trust the results?

We introduce a mathematically verifiable way to understand and check AI model results. Moreover, this methodology allows us to compare and benchmark different AI algorithms.

We utilize domain specific information and expert opinion in this rigorous framework as follows:

- Choose a domain specific machine learning kernel
- Choose a reference data set
- Given black-box AI output on the reference data, calculate the surrogate function (in the reproducing kernel Hilbert space, the unique minimal norm interpolant or best least squares approximant)

- The surrogate function is a weighted sum over the data points of Riesz functions
- From these weights, compare and contrast the relative data point contributions to an AI algorithm outcome
- Domain experts can then interpret and decide whether these attributions make sense
- The attributions of different AI algorithms can be compared and examined for bias
- As new data becomes available and AI algorithms learn, we can monitor this evolution through the surrogates
- The more we understand the transparent surrogate, the more we can trust the black-box AI results

Question 1: How do financial institutions identify and manage risks relating to AI explainability? What barriers or challenges for explainability exist for developing, adopting, and managing AI?

If an explanation is heuristic then a key challenge is that the connection to the AI model is weak and unverifiable. This also opens the possibility of needing an explanation of an explanation.

Beyond explainability, the surrogate analysis that we described earlier is key to rigorous understanding and vetting black box results. Communication of AI model results can begin with analogies and

heuristics, but the explanation is not proof that an algorithm works. These are next followed by a summary analysis of the AI algorithm giving perspective and context. For example, we can highlight the significant differences to alternative approaches. Numerical examples can also be used to demonstrate the sensitivity of results to perturbation, and conclude with recommendations, caveats and applications.

Question 2: How do financial institutions use post-hoc methods to assist in evaluating conceptual soundness? How common are these methods? Are there limitations of these methods (whether to explain an AI approach's overall operation or to explain a specific prediction or categorization)? If so, please provide details on such limitations.

Post hoc statistical analyses evaluate correlation not causation. The latter is what is needed for evaluating conceptual soundness.

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy. For example, in medical imaging why a diagnosis was made is as important as making a correct diagnosis in order to select an appropriate treatment.

A new class of post hoc models to explain a model's predictions, called Shapley additive feature[5] attribution, uses a leave-one-out comparison of features approach to generating a simpler explanation model of a prediction model. This computationally intensive linear method is an example of a surrogate function. By contrast the kernel surrogate functions we have introduced are domain specific, nonlinear and are a more general way to understand predictions beyond marginal contribution.

If an AI algorithm is expensive to compute or unavailable because it is proprietary then one could vet the surrogate in lieu.

Question 3: For which uses of AI is lack of explainability more of a challenge? Please describe those challenges in detail. How do financial institutions account for and manage the varied challenges and risks?

AI applications in finance is relatively new. For example, banks,

[5] Lundberg, S. and Lee, S.,2017,'A Unified Approach to Interpreting Model Predictions', 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

insurers and regulators are just beginning to realize and tackle issues surrounding AI adoption. AI applications are being embraced by those in quantitative finance but the learning curve for financial quants is steep. Moreover, few quant finance curriculums include machine learning. Understanding what an AI algorithm is doing is compounded by dynamic updating.[6]

Kernel surrogates have proven to efficient in machine learning, pattern recognition, signal analysis, scattered data in high dimensions, modeling of geometric transformations, interpolating mesh generation, simulation-based classification, optimal control problems, biomechanical simulations, gas transport problems, density estimation, tuning of hyper parameters, to name a few applications. Scientific industries that use computationally expensive black-box models, including AI algorithms, will naturally have a need to use the more efficient surrogates in the examples mentioned above.

If contextual knowledge makes decision making case by case then the issue of AI explainability is moot. This is why we have trials decided by juries not a computer.

[6] See also our response to Question 8

II. Risks from Broader or More Intensive Data Processing and Usage (Questions 4 and 5)

A key challenge is that AI decisions can be sensitive to minor data perturbations as well as to when regime shifts occur. AI can add significant value but we often don't have intuition as to why. There are several approaches that could be employed to validate the degree of AI algorithm robustness. Perturbation analyses should be adopted so that regulators accept industry use of AI. Part of the reason why regulators have rejected the use of AI is because little attention has been paid to assessing the robustness of algorithms.

Question 4: How do financial institutions using AI manage risks related to data quality and data processing? How, if at all, have control processes or automated data quality routines changed to address the data quality needs of AI? How does risk management for alternative data compare to that of traditional data? Are there any barriers or challenges that data quality and data processing pose for developing adopting, and managing AI? If so, please provide details on those barriers or challenges.

Data may be biased but, in whose favor? This may include the group or certain individuals, as in the tragedy of the commons. Sometimes biased estimates can have lower variance which may be desirable in decision making.

Data perturbation on kernel predictors is an important part of machine learning model validation It is helpful to know if an AI algorithm is sensitive to minor changes in data or what magnitude of data perturbation results in significantly different decision making.

Question 5: Are there specific uses of AI for which alternate data are particularly effective?

The unstructured nature of large nontraditional data sets makes AI exploration of them natural. We are trying to glean insight from these peripheral data sources. But this also cuts two ways, a discerned pattern may not necessarily be predictive. This is exactly why we need to have a sense of what an AI algorithm is doing in order to determine whether we believe a discovered pattern generalizes and is perhaps predictive.

Typically, AI algorithms are parameterized on sets of labeled data points. If the labeling is subjective, without alternative model criteria,

then these sources of nontraditional data necessitate understanding what criteria the AI algorithm came up with for its labeling. This is where kernel surrogates can help us.

The notion of AI solutions which are parsimonious is neither to overfit nor underfit the data in solving a well-articulated problem. Both attributes depend on the utility of the decision maker. One can be rigorous and rational in selecting AI models and making decisions.

Indeed, there is a vast academic literature on decision theory. Nonetheless preference is subjective. In most practical situations we can identify the extremes of overfitting and underfitting data. We discuss this topic further in our response to Question 6.

III. Overfitting (Question 6)

Overfitting an AI algorithm will typically over state in sample accuracy. To rectify this situation institutions should test for themselves out of sample AI predictiveness on consensus benchmark data sets. If an AI algorithm is unavailable to test then one could use its kernel surrogate in lieu. Proof of the pudding, is in the eating!

Question 6: How do financial institutions manage AI risks relating to overfitting? What barriers or challenges, if any, does overfitting pose for developing, adopting, and managing AI? How do financial institutions develop their AI so that it will adapt to new and potentially different populations (outside of the test and training data)?

AI overfitting and underfitting points to the need for parsimonious model selection that is fit for purpose. In most practical situations we can identify the extremes of overfitting and underfitting data. There may however be a range of 'Goldilocks' AI models that meet a decision maker's needs in terms of understanding a problem and being able to extrapolate. If a decision maker is aware of these considerations then a rational model choice can be made that is sufficiently robust and predictive to meet their needs.

There are often multiple mathematical models capable of explaining a given phenomenon. Parsimony is the choice of finding the simplest explanation but not simpler. The rationale being that if a model is too simple then it is likely that it will not faithfully reproduce a phenomenon's behavior while an elaborate complicated model requiring complex specialized assumptions which are less likely to be met and therefore less likely to be predictive. Finding a sweet spot, that

is a faithful robust predictive model, between these two extremes is the purpose of parsimony. Parsimony considerations should also be a major component of human judgment. Between naive and idealistic is practical.

IV. Cybersecurity Risk (Question 7)

As discussed in Question 6, it is one thing to choose faithful and predictive representations of nature that neither overfit nor underfit data. It is an entirely another when an intelligent adversary through a cyber attack is deliberately trying to sabotage your modeling efforts by messing with the data.

Question 7: Have financial institutions identified particular cybersecurity risks or experienced such incidents with respect to AI? If so, what practices are financial institutions using to manage cybersecurity risks related to AI? Please describe any barriers or challenges to the use of AI associated with cybersecurity risks. Are there specific information security or cybersecurity controls that can be applied to AI?

It is well known that AI can learn biased behavior from training data. Perhaps what is less known is that this can be reverse engineered through a cyber attack.

Suppose a malfeasant agent has a predetermined outcome in mind and can generate fake data so that anyone using it to train their AI algorithm will 'learn' this predetermined outcome? For example, there are algorithms that can be deployed (e.g. an algorithm using the Semiparametric Representer Theorem[7]) to systematically create different fake data sets that will lead kernel machine learning towards a predetermined outcome. There is no doubt that there is potential for others to do this.[8]

Emergent Adversarial AI[9] introduces novel game-theoretic cyber security issues to model risk mitigation and parsimonious model

[7] See https://www.cs.mcgill.ca/~dprecup/courses/ML/Lectures/ml-lecture06.pdf for a good discussion on : 1)How to tell if a function is a kernel ,2) SVM regression and 3) SVM classification.

[8] For example, imagine if there was state sponsored distribution of fake data. Scientists, politicians and the public in general do not fully appreciate this potential. In any discipline there should be agreed upon ways to vet and verify the pedigree of training data. Perhaps having benchmark data sets to test algorithms would be a start. Free publicly available data is attractive but has the potential to be malfeasantly tweaked and therefore risk users of the data should not ignore this type of data risk Fake or biased data in machine learning could be dark horse problem.

[9] Goodfellow, McDaniel and Papernot, 2018,'Making Machine Learning Robust Against Adversarial Inputs', Communications of the ACM, vol 61, no 7.

Goodfellow, I. et al. ,2014,'Generative Adversarial Networks' Proc. Neural Information Processing Systems. pp. 2672-2680.

selection. An adversary in a Trojan attack on an AI algorithm may insert mislabeled examples into training data so that the AI algorithm will learn to misclassify the data in a malfeasant way that is advantageous to the adversary. Alternatively, if an adversary knows some of the labels that your AI algorithm is using to classify data then they can create adversarial examples by altering the features of data with a given label in order to make the AI algorithm misclassify these examples without influencing how it was trained.

There are ways[10] to detect AI that is compromised by Trojan horses and adversarial examples but a deeper discussion on this important topic is beyond the scope of this question

[10] Nan Tie, G., (2018), 'Topological Learning', DOI:10.13140/RG.2.2.32873.3440), available at www.researchgate.net.
Nan Tie introduced a topological partition of data wherein we know the average classification of each partition. If we have clean examples of correctly labeled data then after choosing a kernel to use (unknown to our adversaries) we know what each topological partition's average classification should be for this choice of kernel. So, when presented with AI trained on unvetted data if the partition averages are not close to what they should be, then we have potentially detected compromise of our AI algorithm by Trojan horses or adversarial examples. In particular we have a red flag if multiple kernels result in discrepancy. Even if an adversary knows this is what we are doing, it is difficult to create Trojan horses or adversarial examples to avoid detection because they don't know which kernels we will choose to partition the data with. Each additional kernel we use is like adding a tumbler to lock out fake data.

V. Dynamic Updating (Question 8)

AI algorithms such as deep learning neural networks are models whose architecture can adapt and evolve as new data patterns are discovered. For example, an AI pricing model trained on liquid market data may dynamically update in response to a financial crisis. Even though the resultant new pricing may be arbitrage free, the how and why of the dynamic update may not be apparent.

We need to understand what the deep neural network is doing and how it adapted in order to trust its results. Surrogates enable this desired transparency helping us better mitigate model risk.

Question 8: How do financial institutions manage AI risks relating to dynamic updating? Describe any barriers or challenges that may impede the use of AI that involve dynamic updating. How do financial institutions gain an understanding of whether AI approaches producing different outputs over time based on the same inputs are operating as intended?

As explained above, the risk of dynamic updating can be managed by examination of AI algorithm surrogates whose properties are transparent and known.

A dynamic update can work for you or against you. The critical point is that you need to be able to detect the update, understand why the update was made as well as examine the consequences of the update.

Let's take the case where we are using an artificial neural network (ANN) solution. The ANN solution is capable of providing a dynamic update. A key operating concern in using the ANN solution is not detecting that ANN has made a dynamic update. In other words, a key risk is failing to detect that the ANN has made the dynamic update.

Given the surrogate detected an update by ANN, we can now work to understand why ANN made the change. We need to ask questions such as did it detect a change in pattern or did the nature of the data itself change. For example, the change may have taken place due to a movement from a normal to a stressed environment. On the other hand, the change may also have picked up an anomaly.

We need to next examine the consequences and ask if the change is to our benefit. For example, we want to determine if the change still provides us with a parsimonious solution that is fit for purpose.

As a caveat, we should always keep in mind that a surrogate solution does not eliminate model risk. The surrogate is a proxy to the actual AI model and therefore always has an element of model risk.

VI.AI Use by Community Institutions (Question 9)

Community institutions face the challenge that one size does not fit all when applying AI because their experience may not match the data used to calibrate AI. Kernel surrogate explanation of an AI algorithm can be customized to reflect community institutions' views and so mitigate this disconnect.

Question 9: Do community institutions face particular challenges in developing, adopting, and using AI? If so, please provide detail about such challenges. What practices are employed to address those impediments or challenges?

A significant challenge is to address the issue that **various stakeholders have different analytical capabilities**. For example, small scale community institutions generally lack the expertise and resources

necessary to vet and understand AI algorithms that may make them more efficient and fairer.

As mentioned in the introduction, our AI black-box surrogate enables stakeholder interpretation that can be tailored to their views and background. The surrogate provides common ground for users, vendors and regulators to address AI model risk.

VII. Oversight of Third Parties (Question 10)

Institutions need to be confident in and be able to trust in AI developed by third parties. In order for this to happen we first need to understand what an AI algorithm is doing, and then demonstrate its predictiveness, robustness and fairness. The verifiable rigor of causal kernel surrogates enables this, unlike analogies or heuristics.

Question 10: Please describe any particular challenges or impediments financial institutions face in using AI developed or provided by third parties and a description of how financial institutions manage the associated risks. Please provide detail on any challenges or impediments. How do those challenges or impediments vary by financial institution size and complexity?

Adoption and use of vendor proprietary black-box algorithms face the issues of explainability, fairness, monitoring and dynamic updating. Consensus on reference data sets is needed for benchmarking AI results.

If a financial institution selects a vendor then it needs to carefully ask if the vendor solution can answer the question 'Which AI model do I use?'.

We are really asking, 'What problem am I trying to solve?'. If one can clearly articulate the problem to be solved then one can establish the level of uncertainty involved. This is important since identifying the level of uncertainty in a problem guides us towards appropriate types of AI models from vendors and inference to use. Moreover, we then also know what risk measures are appropriate and the potential AI model risk in solving our problem.

The higher the level of uncertainty, the greater the potential for AI model risk. Nonetheless, we are guided by considerations of parsimony in choosing which models are appropriate. Lo and Mueller[11] describe a

[11] Lo, A., and Mueller, M., (March 19, 2010), 'WARNING: Physics Envy May Be Hazardous To Your Wealth!', available at https://arxiv.org/pdf/1003.2688.pdf

continuum of uncertainty. The levels of uncertainty are not meant to be sacrosanct[12] . To reiterate, identifying the rough level of uncertainty inherent in a problem narrows the field of potential AI models to consider, suggests appropriate forms of inference to use and informs us of the potential AI model risk involved.

VIII. Fair Lending (Questions 11 to 15)

Fairness is a consideration in finding a parsimonious AI model solution[13]. Best practice calls for first deciding what you want to accomplish in constructing an AI model, which includes taking fairness issues into consideration. In other words, it is important to write down the question you are trying to answer. You next find a parsimonious solution given that you now know what you want to accomplish. Understanding what an AI algorithm is doing is always desirable. In certain situations, it is critical, for example in medical diagnostics we need to know why a diagnosis was need in order to choose appropriate treatment. Another example is the EU is making it a legal requirement

[12] For example, as with colors in a rainbow, the levels of uncertainty are not meant to be bright lines.

[13] Fehr, E., & Schmidt, K. (2003). Theories of Fairness and Reciprocity: Evidence and Economic Applications. In M. Dewatripont, L. Hansen, & S. Turnovsky (Eds.), Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress (Econometric Society Monographs, pp. 208-257). Cambridge University Press.

that consumers have a right to know why an AI algorithm made a decision that affects them.

Question 11: What techniques are available to facilitate or evaluate the compliance of AI-based credit determination approaches with fair lending laws or mitigate risks of non- compliance? Please explain these techniques and their objectives, limitations of those techniques, and how those techniques relate to fair lending legal requirements.

A prerequisite for evaluation of fair lending is to first know what an AI-based credit determination is doing. Surrogate functions act as if they were the AI algorithm on the data sample. So, if we understand what the transparent surrogate is doing then we have a sense of what the black-box AI algorithm is doing.

As discussed earlier in the Introductory section of our response, proprietary black-box algorithms are sometimes used to make credit decisions. Regulators have used logistic regressions to evaluate such black-box decisions. The use logistic regression here is an example of using a proxy. Our surrogate is a more general, yet finer detailed bespoke evaluation of AI results.

Question 12: What are the risks that AI can be biased and/or result in discrimination on prohibited bases? Are there effective ways to reduce risk of discrimination, whether during development, validation, revision, and/ or use? What are some of the barriers to or limitations of those methods?

Just because AI input - output statistics appear to be biased doesn't necessarily mean the AI algorithm is biased. An AI surrogate may reveal for example the algorithm per se is not biased but the data itself is biased.

A significant barrier or limitation is that it can be difficult to define what is either biased or fair. Fairness[14] depends on societal norms, ethics[15] and benefits to society.[16] For example, AI models which profile certain

[14] Fairness and bias like beauty is in the eye of the beholder.

[15] Useful discussions on AI ethics include:

- LaPlante, A., (2019),'Ethics and Artificial Intelligence in Finance', Global Risk Institute (GRI)

Weldon, D., (April 18 2019)," Understanding the key role of ethics in artificial intelligence, available at:
https://www.dig-in.com/news/understanding-the-key-role-of-ethics-in-artificial-intelligence

[16] Russell S. (Computer Science Division, University of California, Berkeley) CA 94720, Dewey D. (Dept. of Physics & MIT Kavli Institute, Massachusetts Institute of Technology, Cambridge, MA 02139), Tegmark, M. (Oxford University, 16-17 St. Ebbe's str., Oxford OX1 1PT UK), (2015),'Research Priorities for Robust and Beneficial Artificial Intelligence', and Future of Humanity Institute'. AI Magazine,36, No. 4

segments of society may conflict with societal norms. Ethical issues create unique challenges. For example, Google appointed an ethics committee council to deal with ethical issues in AI but it all fell apart[17].

Question 13: To what extent do model risk management principles and practices aid or inhibit evaluations of AI-based credit determination approaches for compliance with fair lending laws?

Current regulation of risk management software is predicated primarily on static code that does not learn and evolve like some AI algorithms. As pointed out earlier, surrogates enable us to understand what a black-box is doing as well as to monitor dynamic updating and manage AI model risk.

Parsimony aside we need to address the fair application of AI based credit models. Statistical models are sometimes used for profiling [18].

---

[17] Bergen, M., Kahn, J. and De Vynck, G., (April 1,2019), Google AI Ethics Council is Falling Apart After a Week", available at: https://www.bloomberg.com/news/articles/2019-04-01/google-s-brand-new-ai-ethics-council-is-already-falling-apart

[18] For example, square root biased sampling has been advocated to screen for terrorism. See Edmonds, D. 'Does Profiling make sense-or is it unfair?', BBC News, available at https://www.bbc.com/news/stories-42328764

The article discusses some of the tradeoffs involved. Once again, we come back to the commons problem.

Models to find fish lead to fishing where there is a high probability of finding fish but there may be serious environmental considerations. AI algorithms already have been found to perpetuate human biases or leverage unseen quirks in data. So, in addition to the parsimonious choice of which models to use, critical assessment of their fairness is important[19].

Question 14: As part of their compliance management systems, financial institutions may conduct fair lending risk assessments by using models designed to evaluate fair lending risks (''fair lending risk assessment models''). What challenges, if any, do financial institutions face when applying internal model risk management principles and practices to the development, validation, or use of fair lending risk assessment models based on AI?

A key challenge is to develop a road map to parsimony when using an AI model to evaluate fair lending risks. For example,

1. Did we build the right AI model?

[19] Fairness is but one of many ethical concerns. Researchers say use of artificial intelligence in medicine raises ethical questions. For example, Stanford researchers discuss the ethical implications of using machine-learning tools in making health care decisions for patients See Hannon,P.,(March 14,2018), 'Researchers-say-use-of-ai-in-medicine-raises-ethical-questions', available at https://med.stanford.edu/news/all-news/2018/03/researchers-say-use-of-ai-in-medicine-raises-ethical-questions.html

- Carefully articulate the problem stakeholders wish to solve
- Identify clean sources of data that can be updated
- Incorporate fair lending risk considerations
- Check the results are fit for purpose

2. Did we build the AI model right?

3. What is AI model risk?

- Quantify the range discrepancy between model results and observed outcomes
- Quantify the impact of AI model risk through stress tests

- Identify the model limitations, including the limitations to incorporate corporate fair lending risk considerations

4. Is the AI model working as anticipated?

- Periodically back test the model
- Update the model parameterization after sufficient new data arrives. An AI model is not necessarily static and may evolve. As previously discussed in Question 8, a critical point in the case of models which provide dynamic updates is that you need to be able to detect the update, understand why an update was made as well as examine the

consequences of the update. Surrogate solutions enable desired transparency

- Review model assumptions. If there is no underlying theory for the AI model results, surrogates of such black boxes can help us monitor potential change and help us better mitigate model risk
- If the model does not work as anticipated then which model do I use

Question 15: The Equal Credit Opportunity Act (ECOA), which is implemented by Regulation B, requires creditors to notify an applicant of the principal reasons for taking adverse action for credit or to provide an applicant a disclosure of the right to request those reasons. What approaches can be used to identify the reasons for taking adverse action on a credit application, when AI is employed? Does regulation B provide sufficient clarity for the statement of reasons for adverse action when AI is used? If not, please describe in detail any opportunities for clarity.

As discussed in Question 11, if a proprietary black-box algorithm is being used to make a credit decision then its surrogate can be used by vendor and applicant alike to openly and fairly assess reasons for adverse action.

All models are calibrated on some training data and then tested on a separate validation data set. AI algorithm calibration is well known to be sensitive to the training data. Moreover, in the absence of an underlying theory of what an AI algorithm is modeling, we need surrogate functions to make sense of what an AI algorithm is doing. In order to calculate a surrogate function, we need a representative sample of input output pairs from the AI algorithm. We recommend that Reg B should explicitly call for making available such samples for third party scrutiny in order to address the fair use of AI algorithms. Kernel surrogates are the scales of justice for the fair use of AI.

IX. Additional Considerations (Questions 16 and 17)

Model risk, the discrepancy between model predictions and actual outcomes, is not a new issue. However, in the case of emerging AI algorithms, two new considerations arise. First, there is still no consensus on a mathematical foundation for AI. Secondly, there is often a black box nature to AI algorithms, in that although we can articulate the steps taken by an algorithm we still may not know what it is doing. Nonetheless these new challenges should not dissuade us from their potential adoption. We just need to be as informed as

possible as we all learn. Technology will always evolve as it has done before.

Question 16: To the extent not already discussed, please identify any additional uses of AI by financial institutions and any risk management challenges or other factors that may impede adoption and use of AI.

Explainable AI based on heuristics or statistical correlations lack explanation of causation. Mathematical AI surrogates on the other hand behave as if they were the AI algorithm on the data sample and have known provable properties to help us understand what an AI black-box is doing, as well as examine what-ifs. If an AI algorithm is expensive to compute or time consuming then its simpler surrogate can be used in lieu.

Stress testing AI models is another challenge. As mentioned earlier, Norvig[20] has recommended stress testing AI algorithms rather than trying to interpret them. Scenario analysis posit 'What if' scenarios that often ignore plausibility, but serve the purpose of empirically examining the effect of inputs on AI model results.

---

[20] Russell, S. Norvig P., (2010),' Artificial Intelligence A Modern Approach, *Third Edition', Contributing writers*: Ernest Davis, Douglas D. Edwards, David Forsyth, Nicholas J. Hay, Jitendra M. Malik, Vibhu Mittal, Mehran Sahami, Sebastian Thrun (Prentice Hall)

Scenarios generated by macroeconomic models themselves are subject to model risk. Scenarios are often based on an arbitrary combination of stress shocks. The danger is that many such combinations may be inconsistent with the basic laws of economics. It is important to examine the chain of events in the scenario and make sure that it makes economic sense. For example, the scenario may violate no-arbitrage conditions.

The potential number of basic stress shocks is enormous. In practice, only a relatively small number of scenarios can be routinely analyzed. This means that the scenarios have to be selected according to the vulnerabilities of the particular portfolio. The usefulness and accuracy of the diagnosis that emerges out of the scenario analysis depends on the judgment and experience of the analysts who design and run these scenarios.

Question 17: To the extent not already discussed, please identify any benefits or risks to financial institutions' customers or prospective customers from the use of AI by those financial institutions. Please provide any suggestions on how to maximize benefits or address any identified risks.

The clinical evaluation of drugs is statistical and we often do not actually understand the underlying mechanism of how a drug works on a body. Nevertheless, we sometimes use things we don't fully understand. We should learn from the current and potential application of AI in other industries such as medicine and apply them to banking.

For example, the FDA has called for discussion of algorithm change protocal because AI algorithms learn and dynamically update[21]. Surrogate functions offer a way to understand what black-box algorithms in medicine are doing and monitor model drift so as to manage the model risk of dynamic updating.

Five recommendations for the adoption of AI by medicine are described below[22]. Comments in parentheses are risk mitigation actions to take that are based on our generic model approach[23]. One doesn't have to

[21] See Artificial Intelligence and Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan, January 2021, FDA US Food and Drug Administration Center for Devices and Radiological Health

[22] O'Reilly,E., (Feb 21, 2019), "Scientists call for rules on evaluating predictive AI in medicine', available at https://www.axios.com/evaluate-predictive-artificial-intelligence-medicine-66a98b6e-7702-4abf-89cc-2430a5e3b3a0.html

[23] Nan Tie, G. and Mark, R.,Sept 2020, 'Parsimony: A Model Risk Paper', Professional Risk Managers' International Association available at https://prmia.org/PRMIAInstitute/Resources/Papers/Parsimony_-_A_Model_Risk_Paper

recreate the wheel. Surrogate functions apply to recommendations 4 and 5.

1. **Meaningful endpoints** that provide clinical benefit from algorithms should be rigorously validated by the FDA (Food and Drugs Administration), such as downstream outcomes like overall survival or clinically relevant metrics like the number of misdiagnoses.

   [Clearly articulate the problem you are trying to solve.]

2. **Appropriate benchmarks** should be determined, similar to the recent example of the FDA approving Viz.AI[24] after it was able to diagnose strokes on computed tomography imaging more rapidly than neuroradiologists.

[Can the algorithm reproduce known results?]

3. **Variable input specifications** should be clarified for all institutions, such as defining inputs for electronic health records so results are reliable across institutions. Algorithms should be trained on data sources from as broadly representative

[24] Viz.AI[24] is a deep-learning algorithm for diagnosing strokes.

populations as possible so they are generalizable across all populations.

[Conduct perturbation and stress analyses to assess algorithm robustness to data variation.]

4. **Guidance on possible interventions** that would be connected to an algorithm's findings to improve patient care should be considered.

   [Can one identity where the algorithm is not working?

   For example, the data is:

   - not representative
   - biased
   - corrupted

   and the algorithm is:

   - sensitive to calibration
   - solving the wrong problem
   - not parsimonious]

5. **Run rigorous audits** after an FDA clearance or approval of a drug in order to check periodically on how the new variables that were introduced via a deep-learning algorithm are performing. The deep learning algorithm may become a less parsimonious choice over time. For instance, regular audits could find the algorithm had a systematic bias against certain groups after being deployed across large populations. This could be tracked in a manner similar to the current FDA Sentinel Initiative program for approved drugs and devices.

[If the initial algorithm choice was parsimonious and fit for purpose then we may find that:

- circumstances can change
- needs can change
- data can change
- theoretical insight can change

In summary, as needs evolve algorithm performance needs to be monitored and intervention anticipated]

D. Concluding comments

Intuitively, a way to think of our surrogate functions is that they stitch together local kernel approximations into a smooth global quilt that is designed to parsimoniously approximate black-box algorithm results. This is why when a black-box is evaluated at a new out of sample data point, the surrogate can tie back to which data points in sample contributed positively and negatively to the outcome.

Domain experts can then interpret and attribute significance to these contributions, in trying to make sense of what the black-box is doing. The surrogate could also reveal the black-box doesn't make sense, or is inconsistent, or even biased. Unlike statistical tests based on correlation, kernel surrogate functions are deterministic and causal. No probability distribution assumptions are needed when using kernel surrogates to understand AI results.

The regulatory community has been increasingly concerned about AI risk models becoming too elaborate[25] (e.g. overly sensitive to the

[25] Curto, C., (May,2013),'Physical and Mathematical Principles of Brain Structures and Function Workshop, available at http://www.personal.psu.edu/cpc16/Curto-whitepaper-2013.pdf Discusses how complicated models impede parsimonious explanation.

embedded assumptions which break down in increasingly more volatile markets).

On the other hand, there may be a risk that the regulatory community discourages the development of more advanced models by moving toward simple standardized models to measure the amount at risk. A simple rule of thumb lacks a coherent rationale for extrapolation, even though it may be based on trial and error experience.

As we have discussed, model building is an iterative learning process that starts with clear articulation of the problem to be solved, followed by identification of the level of uncertainty involved, and so, which types of models and inference are appropriate to use.

AI models may reveal unseen patterns in data but for unknown reasons. Before we throw the baby out with the bath water, let's pause and use surrogates to make sense of what an AI model is doing, so as to assess fairness and bias in decision making. Parsimony considerations then guide our model selection by making the tradeoff between faithfulness and predictiveness. By doing so we choose an AI model that is fit for purpose and understand the model risk involved.

As mentioned in our introduction, our kernel surrogate functions are a parsimonious choice of a data model on a spectrum of model complexity that we can explain. Our kernel surrogate functions fall between simple models like linear regression, and complex models like deep learning networks. Surrogates help us understand what black-box algorithms are doing. This enables assessment of fairness and bias, as well algorithm comparison and benchmarking. Surrogates enable counterfactual what-ifs and stress testing. Moreover, surrogates provide common ground for users, vendors and regulators to systematically vet AI algorithms and mitigate model risk.

E. Authors

Dr. Gary Nan Tie

Dr. Gary Nan Tie, Mu Risk LLC, engages in cross-disciplinary mathematical research, discovering connections across disparate fields to bring new insight in bridging theory with practice. In the beauty of nature there is wisdom. Always the beginner's mind!

Contact Gary at gnt9011@me.com

Dr. Bob Mark

Dr. Bob Mark, Managing Partner at Black Diamond Risk Enterprises, serves on several boards, led Treasury/Trading activities and was a Chief Risk Officer at Tier 1 *banks. He is the Founding Executive Director of the MFE Program at UCLA, co-authored three books on Risk Management and holds an Applied Math PhD. Bob* is a past *GARP Risk Manager of the Year and is a cofounder of PRMIA*

Contact Bob at bobmark@blackdiamondrisk.com