Comments of

ForHumanity[1]
Ryan Carrier, *Executive Director*
Mark Potkewitz, *General Counsel*

In the Matter of

DEP'T OF THE TREASURY Office of the Comptroller of the Currency [Docket ID OCC–2020–0049]
BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM [Docket No. OP–1743]
FEDERAL DEPOSIT INSURANCE CORPORATION RIN 3064–ZA24
BUREAU OF CONSUMER FINANCIAL PROTECTION [Docket No. CFPB–2021–0004]
NATIONAL CREDIT UNION ADMINISTRATION [Docket No. NCUA–2021–0023]

Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning

July 1, 2021

[1] ForHumanity (https://forhumanity.center/) is a 501(c)(3) nonprofit organization dedicated to addressing the Ethics, Bias, Privacy, Trust, and Cybersecurity in artificial intelligence and autonomous systems. ForHumanity uses an open and transparent process that draws from a pool of over 400+ international contributors to construct audit criteria, certification schemes, and educational programs for legal and compliance professionals, educators, auditors, developers, and legislators to mitigate bias, enhance ethics, protect privacy, build trust, improve cybersecurity, and drive accountability and transparency in AI and autonomous systems. ForHumanity works to make AI safe for all people and makes itself available to support government agencies and instrumentalities to manage risk associated with AI and autonomous systems.

Introduction and Summary

Artificial Intelligence (AI), including machine learning, statistical and Bayesian approaches, expert systems, reinforcement learning, and autonomous systems, possesses tremendous potential but presents a concomitant level of risk. Current general approaches to the development of AI systems often fail to account for issues related to ethics, bias, trust, privacy, and cybersecurity in their development, deployment, use, and maintenance. Persons looking to use an AI system should ensure that they understand the specific risks associated with that particular system, including the myriad examples of ethical choice embedded in the design and development of systems. The greater the potential impact on humans, human agency, living creatures, and the environment, the more exhaustive and exacting the scrutiny and analysis that should be placed on those systems.

The financial sector has led the development and adoption of Artificial Intelligence and autonomous systems. Applications ranging from credit analysis to biometric identification of account holders have advanced rapidly and show no signs of abatement. However, the spread of AI tools across various sectors and industries has not increased awareness of the risks associated with these tools. ForHumanity examines the application of AI and autonomous systems when they present a systemic risk to humans, the environment, or societal systems. While most systems are believed to be beneficial, industries must demonstrate that this belief is warranted by building trust in those affected or potentially affected by these systems. The introduction of a robust governance system that embeds human agency, governance, oversight, accountability, and thorough risk mitigations builds trust. Combined with certain advancements in law and regulations in the areas of ethics, bias, privacy, trust, and cybersecurity, industries can, through transparency, accountability, and independent verification, responsibly incorporate AI and autonomous systems into their quotidien systems and practices.

This submission, highlighting the work of 400+ ForHumanity Contributors, explains the risks from these systems and proposes an industry-oriented solution deploying a systemic risk-based approach with transparency and compliance-by-design construction executed throughout the lifecycle of an algorithmic system uniformly across the industry, yet tailored to each individual AI/ML or autonomous system.

Background on Independent Audit

In 1973, the major accounting firms came together and formed The Financial Accounting Standards Board (FASB) which created the Generally Accepted Accounting Principles (GAAP) which still govern financial accounting today. Eventually, the Securities and Exchange Commission, and other extranational regulatory agencies, required adherence to the GAAP (or International Financial Reporting Standards — IFRS) standards for all publicly listed companies. This clarity and uniformity significantly improved the financial world. An infrastructure of trust has been built over the past 50 years because of critical features such as independence, certified practitioners, and third party rules that are compliant with the law and best-practices. ForHumanity has advocated for the adoption of this infrastructure of trust and explained how it can be adapted and adopted for the Governance, Accountability, and Oversight of AI and Autonomous Systems.[2] We support the creation and mandate of Independent Audit of AI Systems (IAAIS).

Role on Independent Audit of AI and Autonomous Systems

IAAIS provides a comprehensive solution grounded in the same fundamental principles as Independent Financial Audit.[3] ForHumanity develops and maintains audit and certification criteria designed for a range of industries and jurisdictions.

The proposed system replicates the distributed oversight, accountability and governance needed for AI and autonomous systems in the same manner as financial audit, through audit and pre-audit service providers. These pre-audit entities will employ certified practitioners to prepare for an eventual independent audit performed by other certified practitioners (Audit). The audit criteria are presented transparently to maximize an entity's ability to achieve compliance. Advancements in systems technology allow many of these processes to be automated for entities such as with the Treadway Commissions' Committee of Sponsoring Organization (COSO) framework for internal risk, audit and controls. The

[2] For more information about Infrastructure of Trust, see Ryan Carrier: #Infrastructureoftrust, Feb. 2021. *Available at:* https://forhumanity.center/blog/auditing-ai-and-autonomous-systems-building-an-infrastructureoftrust

[3] For more information about the taxonomy of IAAIS, see Ryan Carrier & Shea Brown: Taxonomy: AI Audit, Assurance Assessment, Feb. 2021. *Available at:* https://static1.squarespace.com/static/5ff3865d3fe4fe33db92ffdc/t/60329e0a4cfbaa172691f7e6/1613929999802/Taxonomy+of+AI+Audit+%282%29.pdf

result is a fully-integrated, compliance-by-design infrastructure that embeds human agency, transparency, disclosure and compliance from design to decommission.

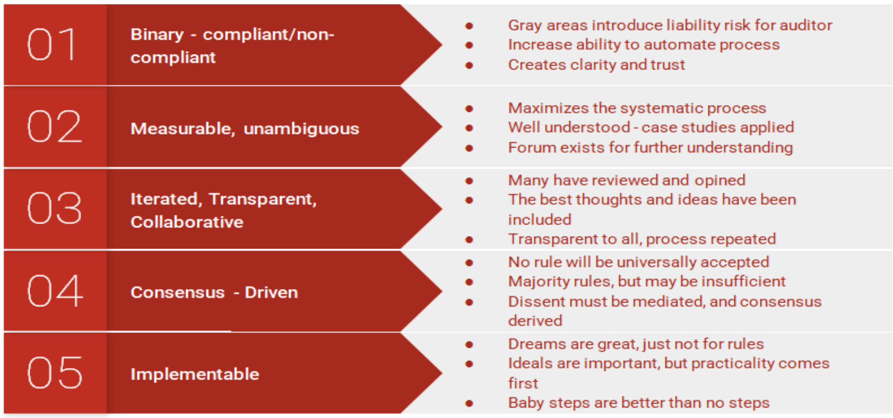
The audit criteria are applied in two vectors: 1) Top-down accountability, governance and oversight 2) laterally, AI system by AI system. The top-down approach creates accountability systems for ethics, bias, privacy, trust, and cybersecurity for the Board of Directors, Chief Executive Officer and Chief Data Officer. Committee structures are required such as Algorithmic Risk, Children's Data Oversight, and Ethics to manage the audit/compliance responsibilities. All of these top-down criteria apply to every AI and every autonomous system in the organization. The system-specific audit criteria is designed to ensure legal and best practice compliance tailored to the specific impact of each system on humans. This comprehensive approach ensures consistency across the organization combined with complete risk management coverage of each unique system.

The creation and maintenance of the Independent Audit of AI Systems is an ongoing and dynamic process. It will continue to be fully transparent to all who choose to participate, provided they join the discussion and participate with decorum. To create each set of audit criteria, ForHumanity engages an international group of experts and seeks points of consensus on its auditable rules. The rules are completely transparent, so when an audit is conducted, compliance is expected.

Independent auditors verify compliance and remain liable for false assurance. Audits must be performed by certified practitioners.

Audit Rules

IAAIS Audit Rules have the following characteristics:



01	Binary - compliant/non-compliant	● Gray areas introduce liability risk for auditor ● Increase ability to automate process ● Creates clarity and trust
02	Measurable, unambiguous	● Maximizes the systematic process ● Well understood - case studies applied ● Forum exists for further understanding
03	Iterated, Transparent, Collaborative	● Many have reviewed and opined ● The best thoughts and ideas have been included ● Transparent to all, process repeated
04	Consensus - Driven	● No rule will be universally accepted ● Majority rules, but may be insufficient ● Dissent must be mediated, and consensus derived
05	Implementable	● Dreams are great, just not for rules ● Ideals are important, but practicality comes first ● Baby steps are better than no steps

These characteristics prove vital for a variety of reasons. Ambiguous audit criteria only encourage auditors to take a more risk-averse approach and presume noncompliance when faced with non-binary choices. Good audit rules must provide the auditor with binary criteria such that certain elements are either compliant or not compliant. The Auditor remains liable for the final report which will either certify compliance or indicate noncompliance. No entity can be certified by an Auditor as partially compliant.

All of these rules must be implementable. Industry can feed into the creation of the rules to ensure that these rules can be followed. In fact, these rules will likely be built into the systems over time for compliance-by-design.

Risks and Pitfalls

ForHumanity is a mission-driven non-profit organization. That mission is *To examine and analyze the downside risks associated with the ubiquitous advance of AI & Automation, to engage in risk mitigation, and ensure the optimal outcome… ForHumanity*. Therefore, we are uniquely positioned to aid the Federal Reserve Board, Comptroller of the Currency, FDIC, CFPB, NCUA, and the industry as a whole to manage these risks. The organizations that design, develop, promote and sell AI/ML tools manage the upside and benefits. Our approach is one of risk control, mitigation, and management. Proper management of downside risks generates better results for everyone. To that end, we have identified five key areas of risk to humans/citizens from applications in the financial industry:

1) Ethics
2) Bias
3) Privacy
4) Trust
5) Cybersecurity

We have developed a transparent, crowdsourced service model for governments, regulators and authorities. We craft audit rules and criteria, submitting them to authorities for approval. ForHumanity will plug into a network of teaching centers to train individual auditors. We license qualified entities to engage in audits or pre-audit compliance partnering with National Accreditation bodies when they exist (e.g., United Kingdom Accreditation Service).

Financial Uses Cases

ForHumanity will develop audit criteria for each of the following specific AI or autonomous system examples, deploying our standard crowdsourcing techniques, inviting industry experts, AI Ethics experts, fintech firms, banks, credit unions, and asset managers to join the crowd and provide criticism, critique, and feedback to the process. This iterative, transparent approach is the refinement process for our criteria, which upon completion is submitted to the appropriate regulatory or legislative body for approval. The following list will be produced in the coming months and may be flagged as a higher or lower priority as regulators may choose.

Audit Criteria Production list (bolded items are already completed):

1) **AML/KYC/suitability**
2) Creditworthiness/credit decisions
3) Flagging unusual transactions
4) Risk Management
5) Cyber risk management
6) Textual Analysis/NLP engines
7) **External Chats/Pricing**/Customer Service/**Chatbots**

Potential Risks Associated with AI

Each of the risks listed below is addressed in detail through crowd-sourced audit criteria and will be submitted for review to governments and regulators.

1) Bias in its many forms
    a) Data Bias[4]
        i) Labelling Bias
        ii) Sample Bias
        iii) Representativeness
        iv) Cognitive Bias
        v) Non-accessibility Bias
        vi) Confirmation Bias and Sunk Cost Bias
2) Legal risk - violation of fairness, anti-discrimination, deceptive practices, data privacy or bias laws
3) Ethical risk[5] - perpetuation of stereotypes and discrimination, failure to abide by a Code of Ethics or anti-discrimination policy, misuse of data, Lack of transparency, Lack of Discloure, Lack of Explainability
4) Model risk - bad validity, bad reliability, poor fit, misalignment with scope, nature, context and purpose,Model concept drift
5) Cybersecurity risk
6) Control and Safety Risk
7) Privacy breaches and failures

[4] For more information about bias in data, see Shea Brown, Ryan Carrier, Merve Hickok and Adam Leon Smith: Bias Mitigation in Datasets found here: https://static1.squarespace.com/static/5ff3865d3fe4fe33db92ffdc/t/60d22a1c1e57e33b08de0cd8/1624386076920/biasInDatasets+%281%29.pdf

[5] For more information about Ethics Committees, see Ryan Carrier, Rise of the Ethics Committee April 2021, found here: https://static1.squarespace.com/static/5ff3865d3fe4fe33db92ffdc/t/60767acef0d59e782d2af79b/1618377424910/The+Rise+of+the+Ethics+Committee.pdf

Responses to Questions

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?***

ForHumanity believes that explainability is a core tenet of governance, accountability and oversight. Some early governance models over similar or related AI systems recognized the importance of explainability and its significance when it comes to how institutions make decisions regarding the use of systems that can impact humans.

While explainability may be difficult or impossible in certain instances of complex neural networks or deep reinforcement learning systems, explainability does not require exhaustingly granular recitations of precise decision pathways. Rather the adversely affected individual must be able to access a better understanding of the decision-making process. If an entity cannot provide such an explanation, that entity may remain vulnerable to accusations of discriminatory or unfair behaviour. Explainability is both a good risk management tool and a fair outcome for humans.

ForHumanity takes a three-fold approach to Explainability. First, our upcoming framework for executing "explainability" establishes sufficient standards requiring differentiation between technical explainability and instances of ethical choice and identifying documentation and transparency criteria and establishing "reasonableness" in the manner in which "explanations" are delivered to the public. These audit criteria require documentation, proof of compliance, and sometimes public disclosures.

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***

ForHumanity advocates for review, compliance, and evaluation throughout the algorithm process from design to decommissioning—as opposed to solely post-hoc reviews. This includes audit compliance on Necessity, Proportionality, Reliability, Validity, and Key Performance Indicators (KPI) design to measure concept drift. These compliance requirements,both technical and non-technical, include extractions of instances of ethical choice where the tensions and tradeoffs can be managed by experts trained in ethical choice. Additionally, we use an iterative bias mitigation process that looks at representativeness or Protected Category Variables and assurance that training-testing/validation data sets are of high quality. However, bias must also be assessed on outcomes. Therefore, if outcomes demonstrate bias,then the bias remediation process must be started again back in the data.

Independent Audit of AI Systems requires the systematic extraction of instances of Ethical Choice during the design and development phase of all algorithmic systems. A common source of failure in models stems from confirmation bias and sunk-cost bias since designers and developers make Ethical Choices in models — choices they are often unaware of and unqualified to adjudicate. Therefore, industry needs trained algorithm ethicists to provide objective feedback on these choices and interface with designers and developers to ensure that necessity, proportionality, reliability, validity, explainability, and KPI design are sufficient for operationalization.

Frequency of KPI testing for concept drift should align to the frequency of learning and/or processing data turnover. Testing and monitoring of any AI system should also occur systematically in operation to ensure consistency with scope, nature, context, and purpose. These ongoing reviews should coincide with internal examinations of training and testing/validation data to ensure sufficient quality, including robust labeling techniques (and testing therein), avoidance of changes in representativeness or outright bias in processing data related to design and technical features. Post-production monitoring of outcomes should be analyzed against fairness metrics.

Initial work on reliability, validity and KPI design around concept drift will provide robust guardrails for the models to remain true to their initial scope, nature, context and purpose. However, failure in any of these key areas might prove terminal to the ethical health of the algorithmic system rendering its decommissioning a more economically efficient or ethically efficient choice over patching, maintenance, or overhaul.

In addition, post-hoc approaches often fail to account for issues such as the need for explainability (see response to Question 1) and the need for diverse inputs in a risk assessment process. These elements, crucial to a proper design and development process, must come into an AI system at its genesis, not its production release. A breakdown or degradation in explainability or a gymnastic *post hoc propter ergo hoc* approach indicate the likely need for the decommissioning of the algorithm.

IAAIS requires diverse inputs and multi-stakeholder feedback in design, development, and pre-operationalization risk assessment in order to identify areas in need of risk mitigation.

3. For which uses of AI is lack of explainability more of a challenge?

A model that is not explainable should not be used in any service where a human, group, or the environment are impacted. For example:

1) financial crimes controls require explainability as a result of evidence gathering, prosecution evidence and the burden of proof;
2) organizational compliance with Regulation B is required under the Equal Credit Opportunity Act ("Reg B");
3) without explainability, identifying data poisoning, model inversion and other Data Entry Point Attacks will be impossible; and,
4) organizations need to balance the tensions and tradeoffs of instances of Ethical Choice in algorithms by the Ethics Committee.

4. How do financial institutions using AI manage risks related to data quality and data processing?

Independent Audit of AI Systems requires entities to operate with a Code of Data Ethics. That Code should be publicly available. A Chief Data Ethics Officer, a Data Ethics Committee (preferred method), a Chief Data Officer, a Data Protection Officer or a Data

Control Committee (preferred method) or any body well trained in Data Ethics must establish or adopt a suitable Code of Data Ethics. The selected person or committee will need to maintain and update the Code of Data Ethics and ensure that staff receive regular training and examination to ensure sufficient knowledge and expertise as it relates to the Code and good data quality which applies "*for both processing data and training data, [and means that] data [is] up-to-date, of high quality, applicable to the algorithm, complete and representative.*"[6]

Data quality includes robust iterative procedures for bias mitigation, that include label testing, sample bias testing, data representativeness, and interactions with the designated individual or body responsible for data ethics within the organization where there are instances of ethical choices, such as benchmarking. Further, the Code of Data Ethics must establish procedures for frequency testing of KPIs around data processing and acquisition to ensure that concept drift has not occurred. The frequency of testing should match to the frequency of new data acquisition.

Data Ethics may be defined as "[t]he branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing, and use), algorithms (including AI, artificial agents, machine learning, and robots), and corresponding practices (including responsible innovation, programming, hacking, and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values)."[7]

Therefore, the Code of Data Ethics must discuss the process for rectifying mistakes and how to test data completeness. For instance, if the model uses synthetic data, the Code must consider when synthetic data may be used, how/when it may be used, the justification or explanation or necessity of its use, and a documentation process for recording its use which includes key decision makers and decision points. The Code must also consider the role of the Ethics Committee in establishing suitable benchmarks and representativeness in training and validation/testing datasets.

[6] This is the definition of Data Quality used in IAAIS.

[7] For more information about data ethics, see Floridi and Taddeo: What is Data Ethics? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2907744

Data minimization should be incorporated into all algorithmic systems, artificial intelligence, and autonomous systems, especially with regards to data unique to any natural person. A Code of Data Ethics should have a process for regular data purges for data that is no longer necessary. Deletion should completely eradicate a link between the Data Subject and the data itself when purged.

An Algorithmic Risk Committee needs to oversee all aspects of best practice compliance, on behalf of the firm's CEO. The CEO is accountable and responsible for assuring all algorithmic systems are compliant with the law, the Code of Ethics and Independent Audit of AI Systems. ForHumanity believes the extension of the Sarbanes-Oxley Act of 2002 in regards to Officer accountability for certifying audit compliance will increase accountability. This Committee shall make sure it considers and examines cognitive bias and non-response bias in each system. Reasonable mitigation should be put in place to manage these risks and minimize residual risk from bias.

5. ***Are there specific uses of AI for which alternative data are particularly effective?***

Since many AI model developers suffer from a lack of access to quality data, they may rely on synthetic or alternative data. The use of synthetic data can carry with it additional risks and serve as a catalyst that can cause a cascading terminal risk reactions such as feedback loops or bias escalation/amplification when improperly used or carelessly incorporated into otherwise healthy models. However, this kind of alternative data can remain incredibly valuable in areas where data droughts, regulations, ethical considerations or statutes prevent regular access to quality data otherwise essential for preparing models and systems. For example, alternative data, such as realistic synthetic data, has proven useful in the space of financial crime.[8] Fraud prediction can use synthetic data to develop better classifiers and lower the risk of unbiased controls. However, care should be taken to ensure that the synthetic data is representative of real data and does not curate further bias.

[8] Lopez-Rojas, Edgar, Barneaud Camille, Advantages of the PaySim Simulator for Improving Financial Fraud Controls. Intelligent Computing-Proceedings of the Computing Conference, 727-736 (London, 2019)

6. How do financial institutions manage AI risks relating to overfitting?

Models have played an important role in finance for decades, and thus model risk management is second nature to financial risk professionals. Therefore, the financial sector, more so than most other sectors, is well-equipped to handle instances of traditional overfitting. Techniques like in-sample/out-of-sample data, stress testing, Monte Carlo simulations and other SR 11.7 risk management techniques have trained financial entities to consider and mitigate this risk.

Given the industry's vigilance on overfitting, we want to focus on the way overfitting can increase adversarial attack vectors through data entry point attacks such as Model Inversion and Membership Inference. ForHumanity defines Model Inversion as the process of reverse-engineering Personal Data, Sensitive Personal Data or Personally Identifiable Information (PII) via the understanding and replication of the algorithmic system and the output. We define Membership Inference as a data mining technique designed to analyze data in order to uncover Personal Data, Sensitive Personal Data or PII. Both attack vectors are widened by overfit models.

Therefore, IAAIS has specific audit criteria which require the Algorithmic Risk Committee to explicitly consider, examine and reasonably mitigate overfitting. Reasonable methods to avoid Model Inversion and Membership Inference attacks can be found in Explanatory note 30[9] and include:

1. Increase training data
2. Reduce (and/or minimize) the size of the model
3. Add dropout/noise during training
4. Regularization
5. Cross-validation
6. Early stopping

7. Have financial institutions identified particular cybersecurity risks or experienced such incidents with respect to AI?

[9] For more information about the structure, governance and operation of Independent Audit of AI Systems see ForHumanity's Independent Audit of AI Systems User Guide v1.1.

The industry may not take proper recognition of the risks associated with bias, or more notably data poisoning attacks. These new forms of adversarial attacks can be broken into two types: 1) Data inputs 2) Training Poisoning. Deepfakes (written, audio or visual) represent a meaningful concern for monitor systems. Each of these forms of attack, often outside the realm of the traditional "cyberattack" represent potential for catastrophic consequences if left unmitigated as they can result in the AI/ML or autonomous system being turned against itself to create either anticipated actions (which can be abused) or outright false steps based upon security protocols which may harm the system or people.

AI/ML/autonomous systems may be a double-edged sword. With systems being designed to increase security, improve safety and enhance autonomy, it is natural to think that these tools will secure the models. However, a number of direct security risks exist, such as Data Entry Point Attacks, unmitigated bias in data sets, and control and safety issues related to what the algorithmic systems may command. These present significant dangers to AI/ML/autonomous systems. In addition, poorly designed systems that fail to incorporate proper ethical choices at key decision points require higher scrutiny. The wrong incident (e.g., a Data Poisoning Attack or security breach resulting in a loss of control) could easily introduce sufficient fear triggering an industry-wide critical re-evaluation which could severely free markets and/or instill doubt in the system.

Therefore, ForHumanity strongly recommends oversight, accountability and governance by design for all implementations of AI/ML/Autonomous systems so that the transparency, disclosure, documentation are constantly reviewed and current. This represents a robust risk management and mitigation process.

8. How do financial institutions manage AI risks relating to dynamic updating?

Independent Audit of AI Systems has a series of requirements related to machine learning and real-time data processing. Criteria compliance requires Key Performance Indicators to manage the risk of concept drift. To avoid sunk cost bias and confirmation bias from designers and developers, the Ethics Committee manages this criterion. These KPIs must have a frequency commensurate to the frequency of data updating to ensure proper and timely tracking of concept drift. The proposed EU AI regulations[10] align with ForHumanity criteria regarding post-production systems monitoring. KPIs for concept drift, examinations of data quality, regular review of data representativeness, labeling, errors, and missing data become necessary maintenance protocols required by IAAIS and require proper maintenance and documentation (and, in some circumstances, disclosure) in order to achieve compliance with IAAIS criteria during an audit.

9. *Do community institutions face particular challenges in developing, adopting, and using AI?*

Community institutions face challenges associated with developing, adopting and using AI. Moreover, small and medium enterprises face a higher relative resource expenditure when complying with government regulations since they often lack full-time General Counsels, or legal and compliance departments. Artificial Intelligence models and systems do allow smaller entities to compete in scope with entities possessing far greater resources through service providers, data aggregators and the availability of the technology and computing power.

The relative IAAIS regulatory compliance requirements will impact SMEs more than large firms which possess the resources to shoulder complex compliance regimes. As a result, the IAAIS framework allows for some marginal flexibility to account for characteristics of most SMEs. However, certain risk areas present too great a potential danger to humans or the environment, and regulators and agencies will need to establish the guidelines and limits.

ForHumanity advocates for the creation of public good, compliance-in-a-box solutions that can be made available to qualified SMEs to enable compliance on par with larger, more resourced firms. This ought to include the assistance of external committee members for great governance, oversight and accountability by maintaining diverse inputs and multistakeholder feedback. Governing bodies that embrace such a model will find greater innovation generated from a wider group of participants and concomitant abatement of

[10] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

risk. Humanity benefits from robust risk management in small firms as well as the world's largest entities and the leverage for this type of approach can be enormous.

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?***

Compliance models are still evolving, but the EU's General Data Protection Regulation provides an excellent model for ensuring compliance across entities deploying third party services by laying out clear and exhaustive rules around contractual roles, disclosures, obligations, and requirements between holders of data (Data Controllers) and those analyzing or otherwise accessing or manipulating that data (Data Processors). Independent Audit of AI Systems allows for data processors, algorithmic and AI service providers to engage with data controllers and contracting parties to ensure compliance for entities seeking compliance with this certification scheme.

A financial services firm looking to hire a service provider must ensure that the service provider is not only aware of the AI regulatory position of the firm but also that the service provider does nothing to alter or jeopardize the obligations, promises, or agreements with respect to the use of AI/ML/autonomous systems. Therefore, a financial services firm must include, in any contract with a service provider,clauses dedicated to specifying audit compliance. It is the responsibility of the compliant entity to ensure that it has acquired all relevant documentation and proof of compliance from the service provider.

AI-in-a-box solutions present additional risks since a user, in this case a firm, can often adjust and customize a pre-trained algorithm taking something that may have once been compliant with IAAIS or some regulatory standard, and adjusted or repurposed it so that it would then fall outside the scope of what can be considered compliant, safe, or ethical.

In these circumstances, ForHumanity recommends the following: 1) The service provider should seek certification and compliance with the original, base model; 2) The service provider should provide built-in compliance mechanisms to enhance the ability of the user to achieve compliance for themselves; 3) The firm should consider training and compliance service assistance from the third party; 4) None of the aforementioned steps eliminates or excuses any compliance responsibilities associated with the firm. In fact, ongoing monitoring, concept drift, degradations of data quality present greater danger in this scenario and should be actively mitigated either by contract or through the establishment of a rigorous model of compliance.

The service providers offering algorithmic systems, artificial intelligence, and autonomous systems should endeavor to provide their clients with information about risk assessments against ethics, bias, privacy, trust and security vulnerabilities throughout their lifecycle with mitigations implemented to reduce risks of non-compliance.

The service providers should also be able to provide documentation that clearly explains how their algorithmic systems, artificial intelligence, and autonomous systems function, for reference by the contracting party. Complexity and size can often be detrimental to the management of these risks and should be followed by precise documentation and comprehensive specificity.

Many Financial Services firms had elected to outsource various aspects of their operations to third party firms as well, therefore any scrutiny required around AI governance needs to be extended to those third party firms. Third party risk management within Financial Services firms, in ForHumanity's view, has not delved into the quality of AI governance in third party firms providing products, solutions and services leveraging ML/AI. Typically, during the procurement process, the due diligence investigations rely upon questionnaires and the occasional deep dive into security matters by the information security teams. The capabilities within Governance, Risk and Compliance functions within Financial Services firms need to evolve and mature rapidly to deal with the extent and quality of AI governance in their service providers to at least the same standards expected within their own organisations. Additionally, security challenges such as model stealing, membership inference, and model inversion,[11] as well as those in our response to Question 7 need to be addressed, as these are embedded and pose significant risks if breached.

The significant gap in third party and vendor risk management relating to AI governance needs to urgently be addressed as more AI/ML/Autonomous systems are adopted and deployed in the supply chain.

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 noncompliance?

Evaluating compliance is the exact purpose of Independent Audit of AI Systems. Fair lending laws and regulatory guidance can be translated into auditable rules (see chart above for an explanation of "auditable"). ForHumanity specializes in crowdsourcing these criteria to make the most comprehensive, thorough, and balanced audit criteria around a particular statute or set of regulations. Once the appropriate regulatory body sanctions the criteria, we make them public and they remain completely transparent. We then ensure proper training and certification standards are established so that only qualified auditors, pre-audit service providers, systems developers, and consultants may license the criteria. ForHumanity works with accreditation agencies, where present, to determine the criteria for qualification or certification. ForHumanity trains and certifies individuals in the criteria, establishing ForHumanity Certified Auditors (FHCA). Auditors are certified against specific schemes, of which there are many varieties based on algorithmic systems, artificial intelligence, and autonomous systems. The Roles and Responsibilities of the members of this ecosystem are explained in this publication.[12] The most important aspect to note is that ForHumanity is a public charity that develops these rules and criteria for submission to a legislature, regulatory body or government agency.

[11] Machine learning security risks in Financial Services, by Adam Leon Smith, Raphaël Clifford, Sarah Carver, Huhan Yang, and Thuy Nguyen

[12] For more information about entity roles and responsibilities - see Ryan Carrier's Infrastructure of Trust in AI - Guide to Entity Roles and Responsibilities, May 2021. https://static1.squarespace.com/static/5ff3865d3fe4fe33db92ffdc/t/60afa2c08b921273acba3a24/1622123209761/Infrastructure+of+Trust+for+AI+-+Guide+to+Entity+Roles+and+Responsibilities.pdf

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?

ForHumanity takes a comprehensive approach to managing bias against legally Protected Category Variables, partly because the complete elimination of bias is impossible. We have extensive criteria in place to consider the myriad ways bias can infiltrate AIs and autonomous systems. From benchmark choices to cognitive bias, from technology barrier bias to sample bias, Independent Audit of AI Systems identified criteria that are binary (compliant/non-compliant), measureable, implementable, and unambiguous. Each of these criteria, when complied with, mitigates bias. The most important part, however, is the governance, oversight and accountability that comes from knowing that compliance will be checked by a third party independent auditor.

An additional critical tool for risk assessment and management is diverse inputs and multi stakeholder feedback. According to Independent Audit of AI Systems, this process occurs at the design and development phase as a means to uncover risks that may not be uncovered by designers and developers. For this criteria, it is critical that the assessment is conducted by a diverse panel of stakeholders designed to represent a broad range of viewpoints, backgrounds, and risk assessment disciplines.

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?

Model risk management principles and specifically model monitoring (SR 11-7) are useful to a certain extent in setting the expectation that the monitoring mechanism shall cover (a) Concept, (b) Process and (c) Computation. However, it does not cover two other essential areas (a) Technology assessment — for cybersecurity or adversarial evaluation and (b) Multi-stakeholder feedback — more like a whistleblower mechanism to report on exceptions or deviation for examination. These feedbacks can come from civil society, regulators and customers.

However, any inhibition may stem from complacency—a hubris that assumes “This is handled.” Understanding the difference between models that sort, filter, classify or establish hierarchy are meaningfully less complex than automated decision-making tools. While many argue that decisions rendered by humans — filled with bias and ranging from ethical to unethical — might be ripe for replacement by machines, they fail to appreciate that these tools have to be managed for bias and built and maintained ethically. Provided that the same rigor is applied by the industry, then ForHumanity believes that SR11-7 and the foundations associated with the regulation will allow for the industry to lead the way in Independent Audit of AI Systems and compliance-by-design models attending to Ethics, Bias, Privacy, Trust, and Cybersecurity.

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?***

ForHumanity has specific governance and oversight designed to ensure accountability for the entire algorithmic process from design to decommissioning, including the extraction of instances of ethical choice, out of the hands of designers and developers. This objective review by a company’s Ethics Committee crosses over necessity, proportionality, reliability, validity, benchmarking, data quality and post production monitoring to ensure avoidance of concept drift. When combined with diverse inputs and multi stakeholder feedback at the design, development and risk assessment stage, plus IAAIS entire bias mitigation processes this collection of auditable procedures provide the public with the most robust process in support of fair lending practices.

15. ***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?***

While Regulation B sufficiently identifies fair results and the prohibition of biased or discriminatory practices in a credit transaction, like most laws, they are typically reactive. A breach or violation occurs and the offending party is retroactively punished and a deterrent effect occurs. There is, however, a more robust process of proactive assurance of compliance with the law, via Independent Audit of AI Systems.

When the law is crafted into auditable criteria, like ForHumanity's GDPR criteria or Regulation B, then Independent Audits of that criteria (approved by regulators) allow for more proactive monitoring that can prevent harms rather than examining them after they occur. Independent audits can result in better application of the law, especially in these more nuanced areas of soft law, reasonable bias mitigation and fairness criteria. A normalized playing field of reasonable, transparent, disclosed and independently verified criteria reduces the risk for firms and the negative impact on humans. The process will document and confirm both reasonableness and fiduciary responsibility are met and exceeded before negative impacts occur. No system is perfect, but a proactive approach with documentation requirements, transparency, normalized criteria and appropriate disclosure will create a virtuous cycle of market-based review, feedback and accountability.

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.***

ForHumanity designs audit criteria to ensure that human agency remains embedded in all algorithmic systems, Artificial Intelligence and autonomous systems. ForHumanity has yet to identify a single use case that justifies an infinite, unbroken operation without the ability to be turned off in some manner or capacity. Therefore, ForHumanity has identified criteria requiring proof that each system can be "turned off." This requirement has more nuance than it may seem. There are academic hypotheses around certain AI, Machine learning, Generative Adversarial Networks (GANs), and autocurricula models that may be able to create workarounds and solutions to avoid being shut off or shut off permanently.[13] Some may perceive this risk as science fiction or Hollywood plot line, but that thinking avoids the

[13] See, for instance, Nick Bostrom's Paperclip thought experiment: Nick Bostrom, *Ethical Issues in Advanced Artificial Intelligence* available at: https://www.nickbostrom.com/ethics/ai.html

fact that these models are designed to learn, based on objectives and often with limited rules or structure. Therefore, to protect humanity, we ensure control with audit criteria. "Off" is one element of control, another element is "avoidance of starting without human intent." "Off" control is meaningless if the system can restart unbeknownst to the organisation. This requirement does not pertain to systems designed to start and stop autonomously by design. This criterion aims to ensure that a system which has been halted by design, remains halted. Control is an important building block to trust. Any system that falls outside of human control will never earn trust and should be considered a systemic risk and decommissioned.

Fairness metrics for outcomes remain elusive for the outputs of algorithmic systems, AI and Autonomous Systems, where a fair outcome is desired. ForHumanity tracks the discussions around fairness. So far, we only use a single quantitative measure of fairness: the Four-Fifths rule[14] which argues that all meaningful selection rates for Protected Category Variables should fall within 80% of the highest selection rate. Absent industry consensus, ForHumanity's framework relies upon the Four-Fifths threshold, accepted by the U.S. Supreme Court regarding employment discrimination claims, as a *de minimus* accepted metric of fairness. Therefore, ForHumanity requires this same threshold on outcomes unless the industry or jurisdiction has declared otherwise.

We endeavor to develop further meaningful measures of fairness. If they become best practice or law, then we are able to build criteria and add these metrics to the IAAIS.

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.***

ForHumanity examines and analyzes the downside risks from these systems, so we will leave the benefits to others. Our final areas of concern lie in second, third and fourth order type of effects. Where unimaginable risks reside today, we need to consider market

[14] 29 CFR Part 1607 - Uniform Guidelines on Employee Selection Procedures (1978).

dynamism and extrapolate future risks when new data are incorporated. As systems' uses increase, systemic risk management techniques such as IAAIS may be ignored or not required which creates even more unmonitored or unmitigated risk. When human agency is excluded from the system, great oversight becomes necessary.

The reliance on algorithmic systems, AI and Autonomous systems can create a systemic risk to the markets not dissimilar to the "Flash Crash of 2010." We can estimate risk from the high correlation of models that "learn" the same thing, at the wrong moment. Built upon massive computing power, correlation can become causation and create a negative feedback wave crashing with tremendous speed pulling many models into convergence upon the same conclusion. If these were market-making models and bids, they could freeze liquidity, and the results could be catastrophic, long before human agency and control could be re-established. Like the proverbial *Stop Button Paradox*, these models might learn improper lessons and circumvent circuit breakers and other market tools designed to allow for pause and reflection on valuations. Creative Risk Management is required to consider what conditions might be precedent to allow for such an event and they should be monitored by authorities and routinely considered by a diverse pool of stakeholders. Traditional Governance, Risk and Compliance governance structures, incorporating AI governance capabilities need to evolve and extend to enable seamless collaboration with technology teams to innovate responsibly when adopting and deploying AI/ML/Autonomous systems through their lifecycle from design to decommissioning.

Algorithmic systems, AI and Autonomous systems must serve humanity first. A profit-first model will result in catastrophic harm to humans without a comprehensive increase in risk management techniques including accountability, governance and oversight that includes independent third party verification.