Pace University

Conceptualization and Instrumentation of Maturity of Responsible AI (MRAI): An Empirical Analysis of the US Banking Industry Credit Lending Practices

John Ratzan

A Dissertation Submitted to
The Faculty of the Lubin School of Business
In partial fulfillment of the requirements for the degree of
Doctor of Professional Studies in Business

New York City
August 2022

Table of Contents

CHAPTER 1 4
INTRODUCTION 4
1.1 Background 4
1.2 Research Purpose 8
Figure 1: Trade-off Framework: 11
1.3 Research Objective 12
Figure 2: Credit Scores 14
1.4 Definition of Terms 15
1.5 Organization of the Study 18
Figure 3: ML Adoption 19
CHAPTER 2 20
RELEVANT LITERATURE and INSTRUMENT DEVELOPMENT 20
2.1 AI background 20
2.1.1 AI capability and functionality 21
2.1.2 Future implications of AI 23
2.2 Credit Risk Model Management 25
2.3 Decision-Making 27
2.4 Responsible AI 28
2.4.1 RAI principles 30
Figure 4: RAI principles analysis 32
2.5 Instrument Development Overview 33
2.5.1 Organizational Commitment & Accountability 34
2.5.2 Transparency, Explainability & Interpretability 36
2.5.3 Fairness and Bias Mitigation 38
2.5.3.1 Regulation 40
2.5.4 Data Management 42
2.5.5 Security 45
2.6 RAI Assessment Frameworks and Toolkits 46
Figure 5: ML Tool usages 47
2.6.1 Capability Maturity Models 47
Figure 6: Gartner AI Maturity Model 48
2.7 Instrument Development Summary 49
2.8 Proxy MRAI Score 49

2.9 Survey Deployment 50
Table 1: Summary of Research papers on principles 52
Table 2: Survey Instrument 59
Table 3: Pre-Validation Survey 61
Table 4: ESG Survey 62
Table 5: Jobin Compendium 63
Table 6: Risk Assessment Frameworks & Fairness Toolkits 64
Table 7: Algorithms 65
CHAPTER 3 66
RESEARCH METHODS 66
3.1 Description 66
3.2 Data 67
3.2.1 Sample Description 67
3.2.1 Data Collection 68
3.2.1.1 MRAI Survey Instrument 68
3.2.1.2 MRAI Proxy Score 68
3.2.1.3 Confidentiality 69
3.3 Validity and Reliability 69
3.3.1 Validity 69
3.3.1.1 Face Validity 69
3.3.1.2 Content Validity 70
3.3.1.3 Construct Validity 70
3.3.2 Reliability 71
3.3.2.1 Internal Consistency Reliability 71
3.3.2.2 Inter-rater Reliability 71
3.4 Summary 72
CHAPTER 4 73
RESULTS 73
4.1 Overview 73
4.2 Instrument Reliability – Pre-Validation analysis 74
4.3 Instrument Reliability - CFA Analysis on Pre-Validation elements 76
4.4 Post-Validation Instrument Reliability 77
4.4.1 Cronbach’s alpha – Organizational Commitment 77
4.4.2 Cronbach’s alpha – Explainability 78
4.4.3 Cronbach’s alpha – Fairness 79

4.4.4 Cronbach's alpha – Data Quality 80
4.4.5 Cronbach's alpha – Security 81
4.4.6 CFA Analysis – MRAI Instrument 82
4.5 Proxy MRAI & Inter-Rater Reliability 83
4.6 Instrument ESG 84
4.7 Validity 85
4.7.1 Multi-Trait Multi-Method (MTMM) 85
4.7.2 Convergent validity – Instrument MRAI with Proxy MRAI Correlation 85
4.7.3 MTMM Matrix: 86
4.7.3.1 Table 8 - MTMM Matrix: 87
4.7.4 Instrument MRAI vs. Proxy MRAI correlation graph: 87
Figure 7: MRAI Instrument vs. MRAI Proxy correlation 88
4.7.5 Discriminant Validity 89
4.7.5.1 Proxy MRAI – Instrument ESG Correlation: 90
4.7.5.2 Instrument MRAI - Sustainalytics Correlation: 91
4.7.5.3 Proxy MRAI - Sustainalytics Correlation: 92
4.7.5.4 Instrument ESG - Sustainalytics Correlation: 93
Figure 8: MRAI Instrument vs. Sustainalytics correlation 94
Figure 9: MRAI Instrument vs. Instrument ESG correlation 95
Table 9 - MTMM Regression Correlation Comparisons: 96
4.8 Capability Maturity Model 96
Figure 10: Bank MRAI instrument survey data in Gartner AI CMM category format. 97
4.9 Normal Distribution Curve of Average Maturity 97
Figure 11: MRAI Instrument Statistical Distribution 98
CHAPTER 5 99
DISCUSSION & CONCLUSIONS 99
5.1 Overview 99
5.2 Theoretical Contributions 99
5.3 Applied Implications 103
5.4 Limitations and Future Research 105
CONCLUSION 107
REFERENCES 108

CHAPTER 1

INTRODUCTION

1.1 Background

Artificial intelligence (AI) (Boden, 2016; Russell & Norvig, 1995; Russell, Norvig, & Intelligence, 2009), defined as endeavoring to simulate the intelligence and rationality of humans (Zackova, 2015), has been ubiquitously adopted by Fortune 500 companies (Davenport & Faccioli, 2019) in their quest to leverage big data insights (Jordan & Mitchell, 2015; Mayer-Schnberger, 2013; Polyzotis, Roy, Whang, & Zinkevich, 2018) to optimize various aspects of their businesses (von Krogh, 2018). The computational power of AI surpasses human capability for mathematical computation by several orders of magnitude (Page, Bain, & Mukhlish, 2018), namely conquering a world chess champion in the widely-referenced Kasparov vs. 'Deep Blue' example (Campbell, Hoane, & Hsu, 2002; Newborn & Newborn, 1997). Competition in the marketplace has increased pressure inside organizations to find new ways to create competitive advantage in terms of speed, efficiency, and cost reduction, including return on investment (ROI) from AI initiatives (Borg, 2021; Digital-Solutions, 2021; Kaya, 2019; Minevich, 2020). AI capabilities have the potential to bring many benefits to firms in terms of boosting revenue and profitability (Biswas, Carson, Chung, Singh, & Thomas, 2020; Kaya, 2019). A few examples leading to increased profitability are leveraging AI to manage banking loan-loss rates, hiring/recruiting screening to mitigate the cost of a bad hire, or lastly reducing the cost of issue resolution from customer service representatives. Each of these functional business capabilities will leverage the AI data, algorithms, and models as well as impact management decision-making in different ways (Boddington, 2017; Boddington, Millican, & Wooldridge, 2017; Boden, 2016; Dignum, 2019). A key management decision that top management teams (TMT) are challenged with is whether to focus on AI efficiency to reduce employees to achieve cost reduction or to leverage AI to redeploy headcount while enhancing capabilities (Davenport & Faccioli, 2019). These companies aggressively strive to innovate in the pursuit of

differentiation, competitive advantage, and profit, while managing the complex web of compliance with corporate social responsibility (CSR) (Dowling & Moran, 2012; Ghadiri, Gond, & Brès, 2015; Satell & Abdel-Magied, 2020) and business regulation (Burt, 2021; MacCarthy, 2020; Pasiouras, Tanna, & Zopounidis, 2009).

There are multiple types of AI that can be deployed to strive for enhancement in productivity. Corporations have largely adopted AI in a 'narrow' form, focused on specific discrete functions (Makarius, Mukherjee, Fox, & Fox, 2020) such as bank lending credit risk management and underwriting decisions (the focus of this study) (Lee & Floridi, 2020; Roszbach, 2003), filtering candidates for hiring (Tambe, Cappelli, & Yakubovich, 2019), and customer service virtual agents (Hunt, 2016). Narrow AI simply means that the AI is fit for a specific purpose, so in the few aforementioned examples, a hiring candidate screening algorithm would never perform a credit underwriting task, and vice versa (Kearns & Roth, 2019). Advanced AI has a specific element named machine learning (ML), of which the key attribute includes a feedback loop in its analysis to optimize its model according to the feedback (Holzinger, Kieseberg, Weippl, & Tjoa, 2018). ML can be deployed in two main methods, which are in a supervised learning fashion or in an unsupervised learning capacity (Jordan & Mitchell, 2015). The main difference between the two methods is that the supervised method of ML contains data labels, so in the case of a simple example of a credit report review, the data would likely consist of the FICO (Fair Isaac Co.) score and the person's name or social security number. Ultimately, the ML program becomes more efficient at predicting and recommending the approval decisions as the machine can process and analyze millions of data permutations scenarios that refine its function (e.g., evaluating the propensity of a loan being repaid) (Jordan & Mitchell, 2015; Roszbach, 2003). In unsupervised learning, the AI/ML is not provided the label definitions for the data, and rather is free to make its own associations, generating AI power, despite being opaque on how the calculations, associations, and predictions from the ML were derived (Adler et al., 2017; Agrawal, Gans, & Goldfarb, 2018; Rai, 2019; Wachter, Mittelstadt, & Russell, 2018).

One of the key aspects of governing AI or determining whether AI is "responsible" or "ethical" is understanding whether the algorithm and model are interpretable, which is called explainable AI (XAI) (Rai, 2019). In models that are explainable, specifically in narrow AI, there is a manageable number of variables that have paths that can be followed in the algorithmic permutations, such that one can feasibly understand the inputs and outputs of the model (Emmert-Streib, Yli-Harja, & Dehmer, 2020; Samek & Müller, 2019). These simpler examples leverage the IF-THEN rules logic of interpretable AI (Boden, 2016). ML also has to ensure to manage data and concept drift provided by the feedback loop to AI (Lu, Zhang, & Lu, 2014). Due to the rapid advancement of AI, some of the newest AI techniques, namely deep and reinforcement learning, are much more powerful in terms of prediction accuracy (Shrestha & Mahmood, 2019). Another aspect of AI/ML is the data (Stoyanovich, Howe, & Jagadish, 2020) that the models leverage to simulate decisions and the role that CDO (Chief Data Officers) have in AI/ML capabilities (Dell, 2020). A number of advances have been made in this component of the technology from data training (Linardatos, Papastefanopoulos, & Kotsiantis, 2020), data ops (Rodriguez, de Araújo, & Mazzara, 2020), knowledge graphs (Tiddi & Schlobach, 2022), to model auditability (Adler et al., 2017) capabilities.

The emerging AI techniques are still in the R&D (research & development) phase and only partially implemented in normal operating processes within companies, and thus may not have models that are explainable, interpretable, and understandable (Emmert-Streib et al., 2020). In addition, the impact and influence of AI on strategic decision-making (Stone et al., 2020), such as M&A decisions, are still in the formative stages due to the complexity, number of variables, and need for 'humans in the loop' (Bussmann, Giudici, Marinelli, & Papenbrock, 2020; Zetzsche, Arner, Buckley, & Tang, 2020). Another example would refer to the 'gut intuition' (Welch, 2001), involved with devising a corporate strategy (Cihon, Schuett, & Baum, 2021) and making strategic decisions (Spangler, 1991; Tambe et al., 2019). There is a growing focus on ensuring that the ML and models that companies deploy have governance of fairness (Holstein, Wortman Vaughan, Daumé, Dudik, & Wallach, 2019), explainability (Linardatos et

al., 2020), security (Papernot & Brain, 2018; Papernot, McDaniel, Sinha, & Wellman, 2016), privacy (de Laat, 2021) and safety (Page et al., 2018) built into them. This governance and monitoring enable humans to understand what artificial intelligence is effectively accomplishing, is the cornerstone of Responsible AI (to be referred to as "RAI" in this study). This study recognizes RAI as artificial intelligence, that can be held accountable to elements of transparency, fairness (bias mitigation), privacy, and security. Many companies have begun to publish RAI principles as organizations begin to embrace the need for consistency in the definition of RAI. (Burkhardt, Hohn, & Wigley, 2019).

Though some RAI principles are published on the topic, many of the firms have different sets of principles, or minimally, principles that are prioritized differently and more importantly principles cannot govern RAI alone (Mittelstadt, 2019). The fact that there are different sets from prominent companies, creates a need and opportunity for a standard definition of RAI (de Laat, 2021). As input for the research for this study, many RAI principles definitions have been identified and published by other researchers (Benjamins, Barbado, & Sierra, 2019; Buhmann & Fieseler, 2021; Clarke, 2019; de Laat, 2021; Eitel-Porter, 2020; EUCommission, 2019; Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; Floridi & Cowls, 2019; Hagendorff, 2020; Jobin, Ienca, & Vayena, 2019; Leslie, 2019; Morley, Floridi, Kinsey, & Elhalal, 2019; Myers & Nejkov, 2020; Rakova, Yang, Cramer, & Chowdhury, 2020; Schiff, Rakova, Ayesh, Fanti, & Lennon, 2020; Siau & Wang, 2020; Zeng, Lu, & Huanfu, 2019) (**Table 1**). In addition to the lack of unified principles that define RAI, there are also various, nascent methods available for assessing maturity (Vakkuri et al., 2021) of Ethical or RAI (Ayling & Chapman, 2021; Bellamy et al., 2018; Clarke, 2019; Digital-Solutions, 2021; Economist, 2020; Fifth-Quadrant, 2021; Mills & Duranton, 2021; Myers & Nejkov, 2020; Saleiro et al., 2019) which set a precedent for the research in this study.

As has been discussed in the context above, there are a few types of AI, and many applications for this technology. Since this study is reviewing the RAI programs specifically regarding banking credit lending, there is a particular focus on the ML of the credit underwriting algorithms, models, and data that are associated with credit lending decisions. With this emphasis on credit lending, there will not be a

detailed explanation of data or use cases related to RNN (recurrent neural network) for NLP (Natural Language Processing) (Baktha & Tripathy, 2017) used in conversational AI, nor CNN (convolutional neural network) for Computer Vision (Face Recognition) (Albawi, Mohammed, & Al-Zawi, 2017; Pouyanfar et al., 2019). In summary, AI is a powerful and rapidly evolving technology that many corporations are adopting, creating a race between the implementation of the evolving capability and the maturity of the governance to ensure safe and fair deployment of AI (Hunter, Sheppard, Karlen, & Baliero, 2018).

1.2 Research Purpose

In the context of this corporate race to rapidly enable these powerful AI capabilities for competitive advantage over rivals, the briskly evolving technology has opened a pandora's box of risk, bias, and privacy issues in managing various operational items and customer interactions (Dignum, 2019). When employing the power of AI, corporations risk introducing biases into automated decision-making, which have attracted regulatory (Candelon, Carlo, Bondt, & Evgeniou, 2021) agencies in a couple of categories related to discrimination, such as the Fair Credit Reporting Act (FCRA), and the Equal Credit Opportunity Act (ECOA), enacted in 1970, 1974 respectively, for lending (ftc.gov, 2020a), and anti-discrimination in employment hiring (Friedman & McCarthy, 2020; Maurer, 2020). Most notably, there is Federal governance, defined in the Supervisory Guidance on Model Risk Management (SR11-7) as the regulation inherent in technology model usage in lending (FederalReserve.gov, 2011b). The bias can pervade the technical capabilities in various ways, including errors, oversights, or unintended consequences with the concern being that AI deployed on a large scale can adversely impact certain groups and individuals (Cowgill, 2019; O'Neil, 2016). Moreover, existing bias in data is exacerbated by the magnitude of computing scale that the AI/ML capability can process, raising concerns with ethics and governance committees within corporations leading many companies to establish RAI programs (Dignum, 2019). AI is a powerful technology that may provide many benefits to corporations in terms of automation, productivity, and cost efficiency; however, there are inherent risks associated with this

transformational capability (Kearns & Roth, 2019). Due to the nature of historical data, AI data training techniques and algorithmic programming methods, bias, fairness and privacy concerns have permeated AI technology deployments (Teichmann, 2019). In some implementations, the AI capabilities are so embedded within the system that they are assumed and almost undetectable (Boddington, 2017). The infusion of AI into this competitive ecosystem has created scenarios in which companies are increasingly required to institute new governance programs, coined as RAI (Dignum, 2019; Rakova et al., 2020) or 'Ethical AI' (Russell, Hauert, Altman, & Veloso, 2015; Siau & Wang, 2020). This rapid advancement and adoption of AI are taxing the capacity of companies to leverage the emerging technology, while ensuring that they adopt the regulatory measures (Samek & Müller, 2019). The emergence of RAI provides a framework for integrating many AI capabilities. RAI reinforces a number of the key Asilomar principles in its description of ethics and values (Asilomar, Institute, & FLI, 2017). RAI focuses on ensuring the ethical, transparent, and accountable use of AI technologies in a manner consistent with user expectations, organizational values and societal laws and norms (Dignum, 2019). Many companies have published RAI principles; however, the principles vary in terms of presence as well as prioritization (de Laat, 2021; Siau & Wang, 2020).

The capabilities of AI become interesting when measuring the maturity of the AI governance in terms of RAI. One of the essential areas to explore, beyond the well-adopted purview of explainable narrow AI, is how companies are managing the emerging AI technology capabilities. Accordingly, companies are managing how humans interact with technology, which is described as HITL (human in the loop) (Rahwan, 2017; Zetzsche et al., 2020). More importantly, this study is interested in how companies are addressing the tension between the quest for accuracy with the innovative AI models and ensuring fairness in society. Corporations remain challenged with competing goals, while corporations invest in AI technology capabilities; they must also invest in RAI to protect against bias and discrimination from a societal fairness perspective (Hategan, Sirghi, Curea-Pitorac, & Hategan, 2018; Lee & Floridi, 2020). Nevertheless, there has been constant tension between profit and societal good in endeavoring to satisfy total return to shareholders (TRS) since the beginning of the modern business era

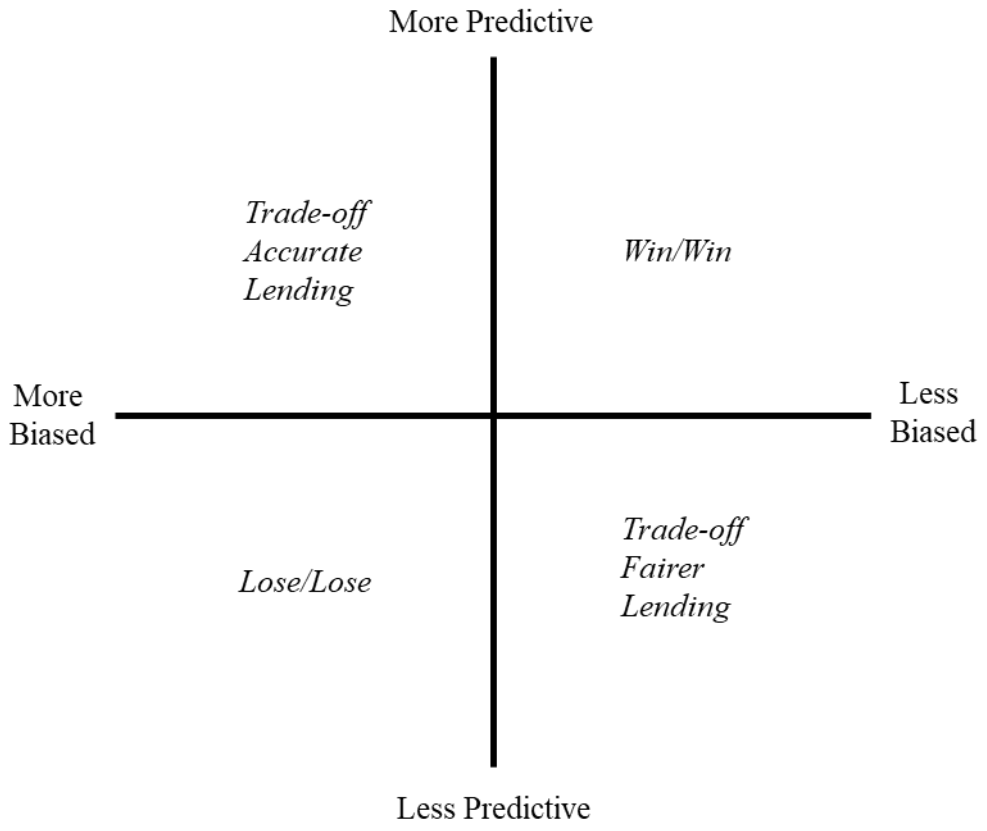
(Friedman, 1970). The prior context established the foundation for reconciling the issue involving the tension of leveraging this powerful virtual machine technology (Boden, 2016) and the need to provide assurance to stakeholders that ethical (Clarke, 2019) considerations are being built into the AI/ML process to protect against harm. Companies need to consider whether they are building algorithms to maximize shareholder profit or to maximize fair distribution of resources to the community (Dignum, 2019). One basic consideration for management is which functions should be strategically deployed for AI/ML, which may introduce societal liability and risk of societal harm. Management may serve to govern the focus on RAI to evaluate tension and ensure that the AI/ML model optimizes profitability and fairness. This is similar in nature to other business regulations, ranging from workplace conditions employee rights, environmental protections, fair treatment in customer rights, to banking supervision for capital ratios, and lastly, safety provisions for food and drug distribution (Pasiouras et al., 2009; Pattberg, 2006; Vogel, 2008). Companies should integrate RAI to balance corporate social responsibility (CSR) exposure and regulatory provisions (Dowling & Moran, 2012; Etzioni & Etzioni, 2017; Ghadiri et al., 2015), in addition to striving for RAI to enhance ROI and related competitive advantage.

When considering decision-making in the TMT, the evolving trade-off and tension between accuracy of optimizing financial performance and supporting fairness to society encompasses AI responsibly in analyzing various types of data (Klein, 2021; O'Neil, 2016; Teichmann, 2019). **Figure 1** depicts a framework on how to assess these trade-offs (Klein, 2021). It is possible that the regulatory mandates are imposed upon corporations may potentially conflicting with profit motives (Askell, Brundage, & Hadfield, 2019; Hategan et al., 2018). While chasing the profitability goals, companies must also adhere to business and societal pressures in terms of compliance and regulation. This creates a scenario in which companies must manage the complexity between introducing AI innovation and complying with regulatory compliance. Ultimately, the goal is to find an optimal point (within the optimal efficiency frontier) (Rao, 1987; Sandmo, 1970) between accuracy in terms of profitability and fairness. In terms of RAI for credit lending, there is an interesting issue to consider regarding investments in AI that pertain to Fraud mitigation; with Fraud investments, there is a clearer ROI that prevents Fraud

loss (Raj & Portia, 2011), but in the case of RAI for lending, where fairness is the goal, there is not a clear ROI present (Digital-Solutions, 2021; Ransbotham, Khodabandeh, Fehling, LaFountain, & Kiron, 2019).

Figure 1: Trade-off Framework:

Below is a trade-off framework that could be leveraged with the Pareto Efficiency Frontier in making decisions on credit loans.



Source: Adapted from Brookings (Klein, 2021)

Due to the nascent nature of AI capability development, there are only scant public references on the investment costs nor any consistency in implementing AI governance programs, which can range from simple auditing to fully mature RAI programs (Adler et al., 2017; Boddington et al., 2017; Koshiyama, Kazim, & Treleaven, 2021; Wall, 2018). Moreover, the financial return derived from implementing RAI is neither well-understood nor widely documented, which creates an opportunity for new research to contribute to RAI. Most of the extant literature focuses on the impact of more narrow

forms of AI (e.g., credit underwriting, automated customer service virtual agents, hiring searches) in terms of profit, efficiency/productivity (Ameen, Tarhini, Reppel, & Anand, 2021; Bussmann et al., 2020; Misheva, Hirsa, Osterrieder, Kulkarni, & Lin, 2021; Roszbach, 2003; Wang, Zhang, Lu, & Yu, 2020; Xu, Shieh, van Esch, & Ling, 2021), specifically the inherent biases (Jobin et al., 2019; O'Neil, 2016), or lastly, describing RAI measurement frameworks (Boddington, 2017; Sen & Ganguly, 2020). There is currently a gap in the literature regarding the optimal balance point between focusing on competitive advantage for financial ROI and managing RAI programs focused on providing transparency, mitigation of bias, and fairness in society. There is also a literature deficit regarding RAI measurement frameworks and instruments insofar as how they are applied to measure the maturity of the RAI programs and the potential correlation with various important corporate attributes. These gaps create a significant challenge as well as an opportunity for corporate TMTs in the context of RAI with their decision-making and optimizing the tension between accuracy and fairness. This leads us to the key research question, which asks 'how does the level of maturity of the RAI program impact the ability to balance the tension and equilibrium point between the quest to attain a competitive advantage delivering healthy ROI (corporate financial performance) and compliance with the key RAI attributes (transparency, ethics, societal fairness)?' This study aims to develop an instrument to enable corporations to measure RAI maturity. Accordingly, we progress into a discussion on the purpose of the research and the contribution to both the RAI theory as well as practical applications for Banks.

1.3 Research Objective

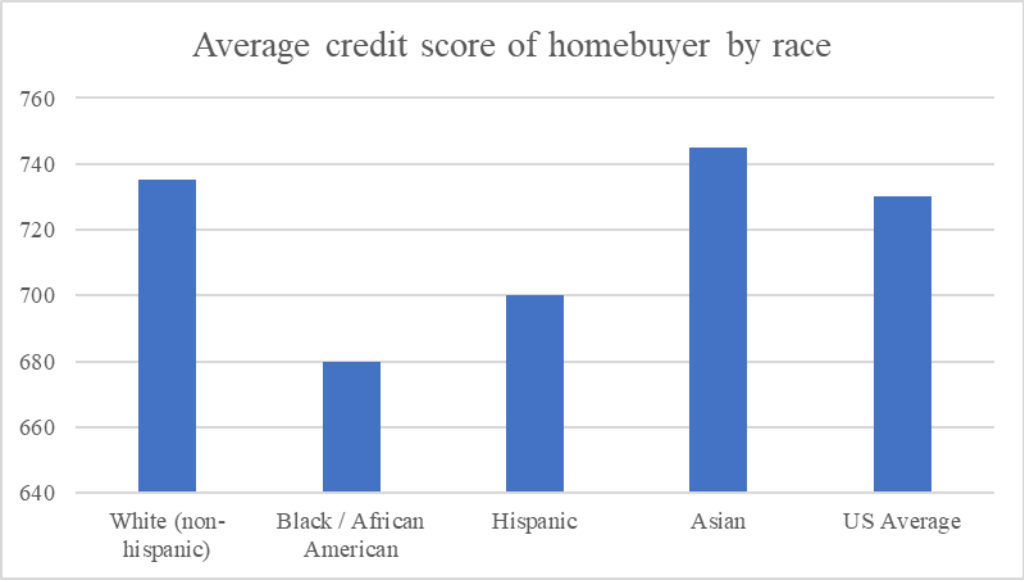
This dissertation proposal will explore the investments, organizational and operating model structure, processes, and capabilities that companies utilize to manage the RAI tension between accuracy in the form of profitability and fairness in credit lending. This study posits that while endeavoring to achieve competitive advantage and associated ROI, firms must possess mature RAI programs that enable them to balance the tension between optimizing the AI for accuracy and supporting fairness to serve an equitable, social purpose. This study will focus on researching the maturity of AI adoption in Bank

credit-lending processes in terms of the RAI framework, and how TMTs are responsibly governing AI in their decision-making. The study will review the maturity of RAI programs that corporations have deployed to manage the powerful risks and benefits of AI in credit lending through the development of a MRAI survey instrument, as this study contends that the maturity of the RAI governance will be directly related to the optimal management of the tension between profitability and fairness. Balancing the cost and benefit of RAI can be measured in a Pareto optimal efficiency frontier curve (Martinez, 2021; Martinez, Bertran, & Sapiro, 2020) in which there are a range of optimal points that are amenable to satisfying the needs of both company profit and societal equity (Askell et al., 2019; Corbett-Davies & Goel, 2018).

There are both theoretical as well as applied contributions to the RAI concepts described within this proposal. On the theoretical side, this study asserts that mature RAI programs will create an environment wherein the quest for the equilibrium point will be prioritized and confer utility for both the company and society thus, the results will be correlated with the best overall financial return for the company. The purpose of theoretical discussion is to provide companies with a strategy on how to develop the best approach to investing in RAI to define the equilibrium point between accuracy and fairness. In terms of how companies can apply the concept described therein, this study aims to explore credit lending practices in Banking and understand the impact of the varying maturity levels of AI governance. In each of these lending examples, by applying the pareto efficiency formula, a company can determine various scenarios that return the best utility function (core-econ & Liebniz, 2021). For example, in the case of a mortgage loan, where the FICO score is a main variable, selecting a high FICO score may deliver the best accuracy, but is potentially associated with false negatives (**Figure 2**), whereas selecting a 'high enough' FICO score will delivery slightly less accuracy, but be more fair (ftc.gov, 2021). There is also a false positive scenario where not lending to a lower FICO, who is actually a good borrower creates a missed opportunity. Exploring and optimizing these scenarios will be the best method to satisfy the utility function for both the company and society (Kaya, 2019; Kearns & Roth, 2019; Koshiyama et al., 2021; Lee & Floridi, 2020). Regression models aspire to generate unbiased results to

achieve an exactness with many samples, and though society would prefer unbiased predictions, they are never based on exact predictions (Agrawal et al., 2018). We believe this study will enhance the competitive advantage for Banks that have invested in developing a mature RAI program, which enables the Bank to find and manage the optimal lending points, as measured by the Pareto efficiency curve to satisfy the tension between accuracy in terms of profit and fairness.

Figure 2: Credit Scores



Source: Brookings – *ValuePenguin report based on Federal Reserve data* (Klein, 2021)

The key contribution to the literature and theoretical framework will build upon current knowledge of AI adoption (Vohra, 2022), RAI principles (Jobin et al., 2019), RAI assessment frameworks (AIEthicist, 2021), existing survey mechanisms (Coates & Martin, 2019), technology in decision-making theory, and AI capabilities by adding a formal measurement instrument for the maturity of the RAI (MRAI) programs. The MRAI score will then allow for future studies to correlate the MRAI with other important corporate attributes in assessing relationships. This study will contribute to the field of RAI by inventing a new measurement instrument and infusing the concept of Pareto optimal multi-objective structural optimization (core-econ & Liebniz, 2021; Rao, 1987; Sandmo, 1970) to examples in

Banks' credit underwriting RAI programs (Dignum, 2019). Another contribution is that some of the prior publications have references to links that may have transitioned or content within the websites may have evolved. The hypothesis is that the Banks that score well in the MRAI survey instrument and have mature RAI programs will be better positioned to manage the tension between accuracy in terms of profit and fairness in credit lending, thereby having a competitive advantage in engaging in valued interaction with customers over firms that have a less mature grasp of RAI.

1.4 Definition of Terms

- AI (artificial intelligence) is defined as the aim to simulate the intelligence and rationality of humans (Zackova, 2015).
- ML (machine learning) is an AI technique defined as an algorithm that contains a data feedback loop to learn from the results of the previous action (Doshi-Velez & Kim, 2017).
- Responsible AI (RAI) is a term that defines a framework for bringing many AI capabilities together. It focuses on ensuring the ethical, transparent, and accountable use of AI technologies in a manner consistent with user expectations, organizational values and societal laws and norms (Dignum, 2019).
- Big Data refers to databases that are built for extensive amounts of data to be analyzed by advanced data processing programs and a key dependency for AI/ML scale (Mayer-Schnberger, 2013).
- TMT refers to the Top Management Team, which are the key decision-makers for the corporation (Certo, Lester, Dalton, & Dalton, 2006).
- CSR (corporate social responsibility) refers to which businesses self-regulate to contribute to societal goals of a philanthropic, activist, or charitable nature by supporting ethically oriented practices.
- ESG (environmental, social, governance) refers to the three central factors in measuring the sustainability and societal impact of an investment in a company or business.
- ROI is a financial term that stands for return on investment which is interested in if there is a positive financial return for the company and in this context for investing in RAI.

- CFP (corporate financial performance) is a term that is related to ROI, but a broader representation of overall financial performance.
- An algorithm is defined as a computer program that comprises a set of instructions for the computer to execute (Kearns & Roth, 2019).
- An AI model is defined as an algorithm coupled with a data set that will be executed by the computer to perform an action, such as make a prediction on a credit application (Koshiyama et al., 2021).
- GOFAI is a term that means the original versions of AI implementation referred to as good old-fashioned AI (Boden, 2016).
- Narrow AI is defined as AI that has a single purpose, such as credit underwriting or playing chess (Page et al., 2018).
- Training data is an extremely large dataset that is used to teach a ML model. For supervised ML models, the training data is labeled. The data used to train unsupervised ML models is not labeled. (Dixon, Li, Sorensen, Thain, & Vasserman, 2018)
- Supervised ML is a type of ML in which the programs rely on defined labels for performing the statistical analysis (Zador, 2019).
- Unsupervised ML is a type of ML that allows for free association of unlabeled data elements, which is potentially a more powerful, but less explainable form of ML (Zador, 2019).
- Explainable AI refers to the amount of understanding or transparency that is available in the AI models. This is important for RAI as the auditability of the models is paramount for RAI compliance and regulation (Samek & Müller, 2019).
- FICO (Fair Isaac Corporation) is a key credit worthiness measure that has historically been used ubiquitously in the banking, payments, and lending industry.
- Data Pipeline is a term that describes the process and toolsets used to automate the movement and transformation of data between a source system and a target repository (Deepa & Ramesh, 2021).
- Data Ops is a term that refers to an organized process to manage the data lifecycle of acquisition, transformation, modeling, mgmt., usage and governance of the data. (Rodriguez et al., 2020)

- Knowledge Graph - A knowledge graph serves as a data structure in which an application stores information, which can be inputted to the knowledge graph through a combination of automated and semi-automated methods. (Tiddi & Schlobach, 2022)
- SHAP (Shapley Values) - SHAP or SHAPley Additive exPlanations is a visualization tool that can be used for making a ML model more explainable by visualizing its output. It can be used for explaining the prediction of any model by computing the contribution of each feature to the prediction. (Främling, Westberg, Jullum, Madhikermi, & Malhi, 2021)
- LIME - The acronym for local interpretable model-agnostic explanations, is a technique that approximates any black box ML model with a local, interpretable model to explain each individual prediction. (Gramegna & Giudici, 2021)
- Pareto Optimization - Pareto efficiency or Pareto optimality is a situation wherein no individual or preference criterion can be better off without making at least one individual or preference criterion worse off without any loss thereof. (Martinez, 2021)
- Security – In ML, security applies to ensuring that the models are secure. (Papernot & Brain, 2018)
- Privacy – In AI, privacy pertains to ensuring that the data are secure, encrypted and not available for public consumption. (Papernot et al., 2016)
- Access privileges – Access privileges are security credentials to be able to access certain components of the computer system, and in this case the ML models. (Wee & Nayak, 2019)
- Encryption – Encryption pertains to the computer function that makes the data unreadable or inaccessible to others that view the data. (Wang, Li, Kuang, Tan, & Li, 2019)
- EDA (Exploratory Data Analysis) – EDA is a technique of visually reviewing the data that will be trained and used in ML models. (Hafen & Critchlow, 2013)
- NLP (Natural Language Processing) is broadly defined as the automatic manipulation of natural language, like speech and text, by software.
- Synthetic Data - Data generated by a computer simulation that is intended to be non-personal and anonymous which is used in training data to preclude bias in the data.

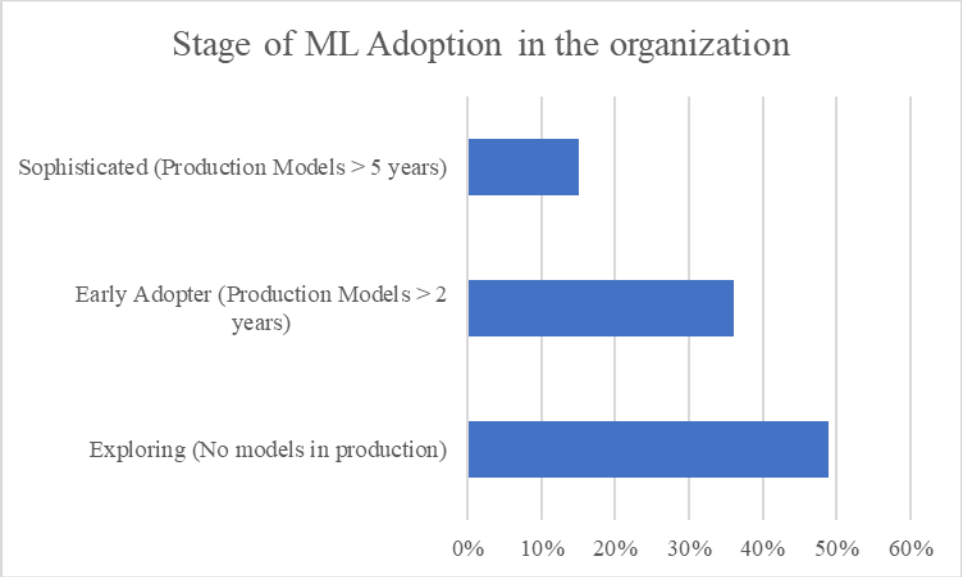
- RNN - A recurrent neural network is a type of artificial neural network adapted to work for time series data or data that involves sequences used in NLP for conversational AI.
- CNN – A class of artificial neural network, primarily applied to analyze visual imagery for computer vision face recognition capabilities.
- GAN - Generative adversarial network is a class of ML frameworks designed to engage two neural networks in a game to ensure data training data is non-biased.

1.5 Organization of the Study

Within this research study, Chapter 1 introduces the definition of AI and ML, demonstrates the power of AI, and highlights the potential risk of ungoverned AI implementations. Due to the potential scale and risk of AI/ML, this governance is even mandated through regulatory law. The study outlines the research question focusing the characteristics of a robust MRAI instrument that will enable Banks to assess the maturity of their RAI programs and related capabilities. The intent is to design the MRAI instrument and then leverage the instrument to collect the data that will both characterize the Banking industry maturity as well as serve as statistical data for the key reliability and validity tests that will be performed. In chapter 2, this study provides extensive detail on the history of AI, and how AI works, along with the various impacts that AI has on corporations and society. This is performed by reviewing a variety of prominent literature on the topic of AI, as well as more specifically on the focus of RAI and related impacts to companies, employees, and customers in the context of credit-lending decisions. Key inputs for the instrument development are identified in terms of common RAI principles, which then comprise the instrument categories. The rationale for the inclusion of the particular categories and attributes is explained in detail. Chapter 3 outlines the methodology, data, and construct validity and reliability tests. The general hypothesis toward the research question and development of the measurement instrument is that Banks are on a spectrum of maturity in terms of adoption of RAI programs and techniques (**Figure 3**) (Lorica, 2018). The new MRAI instrument measures the maturity

through a few key attributes (organizational commitment, transparency, fairness, data management and security), which are detailed indicators for the level of maturity of RAI. The MRAI instrument will be validated against an MRAI proxy score to ensure the validity and internal consistency reliability of the instrument to measure RAI capability. The intention for the data is to go beyond a sample and include an entire population of 50 of the top US Banks. Chapter 4 will discuss results of the statistical analysis using Chronbach alpha, CFA (Confirmatory Factor Analysis), IRR (Interrater reliability) Cohen's kappa, and descriptive statistics. Chapter 5 will provide the discussion and conclusion from the results, as well as managerial implications for the MRAI instrument, and lastly will discuss some of the limitations and areas for future research.

Figure 3: ML Adoption



Source: O'Reilly Survey - https://www.oreilly.com (Lorica, 2018)

CHAPTER 2

RELEVANT LITERATURE and INSTRUMENT DEVELOPMENT

2.1 AI background

AI is categorized as a GPT (general purpose technology), which can bring a broad range of impacts to society (Brynjolfsson, Rock, & Syverson, 2017). Originating in the 1950's by the novel invention and study of McCarthy (McCarthy, Minsky, Rochester, & Shannon, 1955), Minsky (Minsky, 1961), Turing (Turing, 1950), and Simon (Simon, 1973) to name a few, AI has exhibited dramatic optimism and endured cold winters since its inception (Buchanan, 2006; Campbell, 2019; Pan, 2016). AI has progressed over the past few decades from simple computer task automation to advanced ML, entering into areas of face recognition, with Computer Vision (Feng, Jiang, Yang, Du, & Li, 2019) and conversational AI, not to mention speech recognition with NLP (Natural Language Language) (Zaib, Sheng, & Emma Zhang, 2020) with the potential to eventually become superior to the human brain with AGI (Artificial General Intelligence) (Tegmark, 2017). AI is a prediction technology, and predictions are inputs into decision-making (Agrawal et al., 2018). A thought comparing computers and humans is that computers have two structural advantages over the human brain in that first, the computational power is far greater, but second that computers are able to be linked together, which currently is not feasible with human brains truly defining bounded rationality (Kanaan, 2020; Simon, 1955). In the modern age, many high tech companies (e.g., Amazon, Google, Microsoft, Apple, Facebook) have reached amazing valuations based on leveraging technology (India, 2019). Governments are also keen on the power of AI and developing national strategies to ensure they understand AI and can leverage this feature in the geo-political theatre (Brynjolfsson et al., 2017; Kanaan, 2020; NSTC, 2016). There is a continuum in the different types of AI that companies deploy for various business objectives. Essentially, the technology began with classic or symbolic AI (helping to improve current business practices) known as GOFAI (good old-fashioned AI) (Boden, 2016). A more advanced form of AI is augmented narrow AI (focused AI on discrete tasks enabling new capabilities) (Page et al., 2018). Lastly, a recently emerging form of AI

is called autonomous AI including deep and reinforcement learning techniques (Boden, 2016; Jordan & Mitchell, 2015) where the AI involves the AGI (artificial general intelligence) (Brynjolfsson & McAfee, 2016) cognitive progression to enable human-like independent decision-making capabilities) (Davenport & Kirby, 2016; Garbuio & Lin, 2018).

If continuous advancement in technology has always been a focus for competitive advantage, then why are the AI capabilities now at such an inflection point (Chander, Srinivasan, Chelian, Wang, & Uchino, 2018)? There are a few factors that have been evolving over the past few decades since the inception of AI. These components reduce the cost of computing in accordance with Moore's Law (Moore, 1965), and the generation of data as well as mining capability inherent in big data (Chen, Chiang, & Storey, 2012; Duan, Edwards, & Dwivedi, 2019). AI/ML advancement (Jordan & Mitchell, 2015), and cloud-based platforms reduce capital investment needs for computer processing power (von Krogh, 2018). The ubiquitous nature of technology as a fundamental tenet of competitive advantage that has mandated significantly more investment in 2021; the aggregate industry cost of IT investment is ~$3.8 trillion dollars (Lovelock, 2020).

2.1.1 AI capability and functionality

The more recent innovative capability of ML elevates AI through the ML technique, where AI begins to self-optimize or 'learn' outside of human intervention (Baum & Haveman, 2020). This 'learning' does not happen in a vacuum, however, and at least for the contemporary era, AI/ML must have a human architect to program the algorithm models, manage the data, and provide the models training data to teach the program to self-optimize (Jordan & Mitchell, 2015). This 'artificial' intelligence in based on instant data lookup and analysis capabilities that allow for superhuman computational prowess (Davenport & Kirby, 2016). Society has experienced computers using AI to accomplish amazing feats, such as conquer a world champion chess grandmaster (Garry Kasparov) (Newborn & Newborn, 1997), beat a Go game champion (Lee Sedol), and even win against Jeopardy game champions. Computers have quickly exceeded human 'bounded rationality' (Simon, 1947) as

demonstrated by narrow AI (Page et al., 2018) capabilities such as route mathematical and data processing capacity. The historical benchmark for artificial intelligence was first coined by Turing, which states that the Turing Test is a conversation with an AI in which a human is not able to decipher the difference between a human and a machine (Turing, 1950). Corporations are acutely focused on existing AI capabilities, which are described as narrow AI (task level automation) (Page et al., 2018). Evolving AI such as Turing like AGI (Artificial general intelligence (advanced cognition and reasoning) (Turing, 1950), and superhuman intelligence (self-aware conscious systems) (Bostrom, 2014; Kaplan & Haenlein, 2019) are seemingly decades away from transforming from science fiction into reality (Tegmark, 2017).

Technology is a driver of business model change (Garbuio & Lin, 2018). AI is fundamentally different and more powerful than most technologies in that it is pervasive in terms of the various impacts it can make on organizations and society using abductive (Bamberger, 2018) reasoning. In this case, abductive reasoning is important as it deals with uncertainty in the outcome due to incomplete observations, whereas deductive reasoning assumes there is a guaranteed conclusion; alternatively, inductive reasoning assume that a conclusion is likely. AI/ML has had a profound impact on the strategy, process, and structure of organizations (Baum & Haveman, 2020). There are several components to consider regarding the impact and governance of AI capabilities, which are used in various corporate functions, organizational structure, data management techniques, decision-making, risk tolerance, and the rise of RAI governance programs (Rakova et al., 2020).

AI can manifest itself in myriad ways for corporations from predictive algorithms for propensity to pay mortgage loans (Deng & Gabriel, 2006; Murawski, 2019; Rodriguez, 2020) to fighting fraud (Bolton & Hand, 2002) and preventing adversarial security breaches (Robertson et al., 2016) to optimizing employee hiring (Tambe et al., 2019) to automated customer service virtual agents (Adam, Wessel, & Benlian, 2020). An example would be the loan process and credit decisioning, where a FICO score is analyzed along with some other data, and an automatic decision is made by the AI algorithm (Wachter et al., 2018). AI includes several tools, techniques, and algorithms, and is expanding quickly

(Jarrahi, 2018). Some of the terminology of the components is expressed in terms of conversational AI and natural language processing (NLP), which is used by call-center virtual assistants, and ML, which is used in prediction algorithms (Agrawal, Gans, & Goldfarb, 2019); also, computer machine vision (MV), which is used in the analysis of images (Russell et al., 2009).

Corporations serious about employing AI must navigate through an 'AI Life Cycle' (Tambe et al., 2019) to deploy this capability. This involves conducting ML pipeline management, which consists of generating and collecting data, training the ML through data, and then applying the various outputs of the ML to decision making (Hunter, Sheppard, Karlen, & Balieiro, 2018). In consideration of the societal, ethical, and legal (Cath, 2018) implications of implementing AI, it is essential for the architects of the algorithmic programs to be aware of and be responsible for the outcomes (Dignum, 2019). It is also important to understand the decision-making models of research (deductive, inductive, abductive) in applying AI/ML to strategic decision making (Leavitt, Schabram, Hariharan, & Barnes, 2019).

2.1.2 Future implications of AI

There is no lack of trepidation among humans with introducing such powerful technology capabilities (Tegmark, 2017). In employing this powerful AI technology, there are significant concerns that have been raised in the context of the impact on society which range from workforce displacement (Kavanagh, 2019), privacy concerns (Holzinger et al., 2018), potential bias in decision-making (Teichmann, 2019), and most importantly, loss of control over the AI and associated robots (Tegmark, 2017). In fact, there are many references to the industrial revolution and the Luddites with concerns about technological unemployment (Brynjolfsson & McAfee, 2011, 2012). However, at least in 2022, AI can't even begin to compare with innate capabilities that babies have of recognizing images and learning naturally, which is known as Moravec's paradox (Moravec, 1988). In a management example, this is referred to as the 'smiley line' graph, where operational items are far superior with computer machines, but intuition and social interaction remain currently dominated by humans (Ferràs-Hernández, 2017). So far, AI has performed very well with a category of challenges named WSP (well-structured problem), and

the next domain to conquer will be in the ISP (ill-structured problem) (Simon, 1973). Another important set of pairing is Narrow (weak) AI equated with Supervised ML, AGI associated with (strong) unsupervised learning and Super Intelligence coupled with unsupervised and reinforcement or deep learning (Pouyanfar et al., 2019; Shrestha & Mahmood, 2019). Though GPT stands for general purpose technology, the term GPT shouldn't be confused with more recent innovations also named GPT provided by OpenAI's GPT-1, GPT-2, GPT-3 (Generative Pre-trained Transformer) API (Application Programming Interface) (OPENAI, 2021) for NLP (Natural Language Processing) (Zaib et al., 2020).

Advancing to a case of a more holistic strategic decision-making scenario (Stone et al., 2020), such M&A (Mergers & Acquisitions) where one questions if company A should acquire company B, a simulation (an AI program) would have to be performed to incorporate several variables that lead to overall scenario-specific outcomes. (Davenport & Ronanki, 2018; Hunter, Sheppard, Karlen, & Balieiro, 2018; Spy, 2018). In addition, leveraging Monte-Carlo simulations as a supplementary tool for strategy is beginning to exhibit a superior foundation for strategic decisions (Silver & Tesauro, 2009). This type of AI used on strategic decision-making, which involves judgement, is defined as the ability to analyze tradeoffs and make decisions on payoffs (Agrawal et al., 2019). This would involve building complex models representing scenarios and trying to simulate competitive environments (Ferràs-Hernández, 2017; Spangler, 1991) where the VUCA (volatility, uncertainty, complexity, and ambiguity) is high. Another perspective on AI is that the optimal strategy is not to compete with machines and rather leverage their natural strengths to collaborate with human ingenuity (Daugherty & Wilson, 2018, 2022). In fact, Chess is not singularly dominated by supercomputers, and rather the world champion of Chess is a combined 'centaur team' of humans and computers (Brynjolfsson & McAfee, 2012; Follett, 2019; Harari, 2017). (Garbuio & Lin, 2018).

There is a key question to discuss within the topic, which is whether a non-sentient intelligence can be a decision-maker in traditional management (Balasubramanian, Ye, & Xu, 2019). Another concept treats the AI system the same as you would a human that continues to evolve and necessitates

evolving management techniques for productivity, safety and purpose (Brynjolfsson & McAfee, 2016; Tegmark, 2017; Teichmann, 2019). There is also a very real and not too distant potential that AI capability may enhance to a level where it may replace many management decisions (Ferràs-Hernández, 2017). Another key question to explore is if AI will advance to a degree that it can replicate and exceed other forms of cognition, such as emotion, and intuition such that it could actually lead humans in a company (Ferràs-Hernández, 2017). In order to truly begin to make management decisions, AI will have to evolve and behave more like a sentient being (Balasubramanian et al., 2019; Zeng et al., 2019).

Organizational and 'combinatorial innovation' through new structures, processes and business models will be key to leveraging the power of AI (Brynjolfsson & McAfee, 2012). There is relative consensus that this type of human replication in thinking is decades away (circa 2022), and some even would conjecture that the idea is ultimately science fiction (Arnold & Scheutz, 2018). Discussions on topics such as the 'Singularity' and 'Spiritual Machines' by Kurzweil (Kurzweil, 1999, 2006) and Transhumanism (Bostrom, 2005) where the human consciousness is actually uploaded into a computer are presumably many years off, but not impossible. Discontinuity theorists paint dystopian scenarios of AI superiority and human decline while continuity theorists believe that the human-machine interaction will be sustained for the foreseeable future (Shestakofsky, 2017).

2.2 Credit Risk Model Management

Banks earn a significant amount of their revenue at their highest margins through credit lending in various well-known financial products, such as underwriting mortgages, providing auto loans, and issuing credit cards (Abedifar, Molyneux, & Tarazi, 2018). Each of the credit lending underwriting decisions effectively assesses the risk that the individual borrowing the money will re-pay the loan (Fatemi & Fooladi, 2006; Trivedi, 2020). A credit score's usefulness for a corporate lender depends on the ability to predict the borrower's performance (Langenbucher, 2020). Credit risk can be defined as the risk of potential loss to the Bank if a borrower fails to meet its obligations (interest, principal amounts) and is the single biggest risk for a Bank (Leo, Sharma, & Maddulety, 2019). Credit risk techniques have

been evolving for the past 50 years beginning with ration analysis and MDA (multiple discriminant analysis) (Altman, 1968). Enhanced and effective risk management has the ability to provide financial institutions with savings on the order of hundreds of millions of dollars annually (Butaru et al., 2016). Technology, with the advancement of AI, has enabled Banks to do more analytics on each potential borrower and at near instantaneous speed (Misheva et al., 2021). The use of ML in the credit lending process is a function of the algorithm based on the data and the model predicting various aspects about the loan and the borrower's propensity to pay back the loan (Truby, Brown, & Dahdal, 2020). One of the key data elements used in this analysis is the credit score largely managed historically by FICO and is used by the main credit agencies (Langenbucher, 2020). The use of FICO and other data elements has raised the alarms of many potential borrowers as well as associated groups that may be disadvantaged due to some of the data that the algorithm uses to perform the underwriting analysis (Cheng, Varshney, & Liu, 2021). However, there still is not consensus on an overall best statistical method for credit-scoring modeling (Figini, Bonelli, & Giovannini, 2017).

There are a variety of research papers that detail the use of credit risk models in different use case scenarios. An innovative use of data was demonstrated by leveraging commercial loan data and then cross-referencing with AI/ML to build better interpretable risk models to facilitate lending decisions with consumer loans (Khandani, Kim, & Lo, 2010). In a different example, with Italian commercial manufacturing companies, the result of comparing traditional statistics to an Artificial Neural Network (ANN) led to mixed results with varied strengths and weaknesses for each approach (Pacelli & Azzollini, 2011). Competition is driving banks to employ best credit risk capabilities as evidenced in a survey of German savings banks (Bülbül, Hakenes, & Lambert, 2019). In another scenario, there seems to be a significant advantage of ML techniques over traditional statistical regression methods cited in an example on bankruptcy prediction (Barboza, Kimura, & Altman, 2017). In a different study in the commercial banking space, the use of ML forecasts saved a potential range of costs savings of 6% - 25% of total losses (Khandani et al., 2010). Another example leveraging German credit data, again found that ML techniques (using Chi-Square) were beneficial in terms of an improvement in the credit scoring prediction

(Trivedi, 2020). In terms of a comparison of the technical aspects ML, a paper that compared various ML technique on Lending Club loan data found that the random forest algorithm was the most effective (Zhu, Qiu, Ergua, Yinga, & Liu, 2019). **Table 7** depicts a summary of the popular ML algorithms for supervised and unsupervised AI/ML for model risk management capabilities and audits is expected to increase using ML to help govern ML. In this case, leveraging unsupervised learning algorithms to help in the model validation process with monitoring of internal and regulatory stress-testing models, ultimately assessing whether those models are experiencing symptoms of drift (Tammenga, 2020). Banks will continue to invest in AI/ML technology to enhance their ability to perfect the risk balance between maximizing loan profitability and minimizing loan defaults (de Castro Vieira, Barboza, Sobreiro, & Kimura, 2019).

2.3 Decision-Making

Traditional management and strategic decision-making accentuating the role of emotional intelligence (Goleman, 1995), intuition (Welch, 2001), and bureaucratic organizational politics (Weber, 1905) have reigned for the past century of business leadership. Technology has impacted leadership and management significantly over the past 80 years completely transforming the world of business (Liker, Haddad, & Karlin, 1999). AI contends to make an even bigger impact on management and how leadership decisions are made (Ferràs-Hernández, 2017). What is unclear is how much of the decision-making in these cases is being performed without any human intervention and what are the implications? AI is significantly employed in tactical decision making as well as beginning to become a more critical aspect of strategic decision-making as the overall capability matures. Advanced use of AI in decision-making is known as ‘algorithmic management’ (Christian & Griffiths, 2016; Tambe et al., 2019). This raises a key point of comparing effectiveness of shifting to rely on data from AI algorithms for strategic decisions vs. traditional intuition and experience-based decision-making (Simon, 1987). Within some of the decision-making processes in corporations today, the Pareto Efficiency frontier is utilized where there are multiple stakeholders or variables that need to be solved for (Martinez et al., 2020). Decision makers

involved with the credit lending process rely on this analysis tool to balance profitability with fairness in the credit lending decisions (mortgage, auto, credit cards) (Zhou et al., 2021). There is also an important aspect of human involvement called HITL (Human in the Loop) (Rahwan, 2017), in which the human will either derive the information from the AI output and make a decision or minimally review the decision of the AI model and provide an approval.

2.4 Responsible AI

RAI is about being responsible for the power of AI (Dignum, 2019). RAI (also referred to as AI Ethics) (Coeckelbergh, 2020) is acutely relevant when the delegation of human decision-making to AI occurs (von Krogh, 2018). This research studies the specific application of decision-making in the context of Banking credit-lending decisions (Wang et al., 2020). There is a distinction between RAI and AI Ethics, however, where ethics refers more to morals and values, whereas RAI is the practical application of the guardrails to mitigate bias and preserve privacy (Dignum, 2019). This focus on social organizational justice and fairness in RAI is a key focus of this study in researching whether RAI is associated with competitive advantage and higher financial performance for corporations. Examples of where RAI is most critically needed is in cases where there is a potential for bias or error in the data, model (algorithms) programs, or training techniques for the models, and ensuring the proper controls are in place for governing the capabilities of the technology (Cihon et al., 2021; Kavanagh, 2019; Žliobaitė, 2017). There may be cases where bias exists, but the Bank can offer some counterfactual explanatory evidence to support rejecting the loan (Ghosh, Prasad, & Pallail, 2021; Morley et al., 2019). RAI is a framework for managing these key elements collaboratively and should meet specific conditions of accountability, responsibility, transparency (ART) (de Laat, 2021; Dignum, 2019). It focuses on ensuring the ethical, transparent and accountable use of AI in a manner consistent with fairness to stakeholders, as well as upholding organizational values and societal expectations (Burkhardt et al., 2019; Žliobaitė, 2017).

This study defines RAI as the ability to implement AI/ML models that can transparently explain the data inputs and predicted recommendation outputs of the models such that fairness, in terms of mitigation of bias and harm, is confirmed.

RAI can help prevent the use of biased data and algorithms (Gramegna & Giudici, 2021) and ensure that the actions and automated decisions resulting from the models are interpretable and explainable (Doshi-Velez & Kim, 2017; Linardatos et al., 2020; Samek & Müller, 2019). In making the capabilities of AI explainable and transparent, maintaining user trust and individual privacy is critical (Wachter et al., 2018). By providing clarity into the governance of the use of AI components, RAI allows organizations to innovate and realize the transformative potential of AI (Coeckelbergh, 2020; Hunter, Sheppard, Karlen, & Balieiro, 2018; Rakova et al., 2020). A prominent example of the importance of RAI is in reference to DeepMind acquisition by Google, where part of the contractual agreement for the transaction was to deploy a clear Ethical AI framework (Kearns & Roth, 2019).

There are a few key elements that need to be considered when deploying AI to ensure that the organization can maintain control, demonstrate fairness and responsibility, and maintain accountability of the capabilities (Boddington, 2017; Dignum, 2019). Ultimately the goal is to be able to establish an AI that is explainable, which means that the system can provide a rationale for the decisions and characterize the elements of the decision-making process (Ashoori & Weisz, 2019; Holzinger et al., 2018). Responsibility in AI is also an issue of regulation and legislation as it ultimately relates to liability (Burt, 2021; Dignum, 2019; Truby et al., 2020; Wall, 2018). Additionally, the system should be able to provide predictions into the output of the models and be able to verify accuracy in those predictions in a pilot capacity before scaling into a big data set (Adler et al., 2017; Liebergen, 2021).

RAI must include participation to ensure that the AI systems will meet their societal and ethical principles (Dignum, 2019). The foundation for RAI is the framework to be able to govern the various components of the AI capability (Floridi & Cowls, 2019; Rakova et al., 2020; Wright & Schultz, 2018). A critical component of the framework is defining guidelines around the design of the AI system as a

baseline concept (Jobin et al., 2019). The design must engender trust and provide transparency and explainability as to the rationale for why the results of the data, model, and algorithm were produced as such (Holzinger et al., 2018; Taddeo, 2017). These results should be able to be examined for bias and provide evidence of fairness (Saleiro, Stevens, Anisfeld, & Ghani, 2018; Teichmann, 2019). Another key element of RAI is the care in the data selection, collection, and management such that bias does not creep into the overall system (Emmert-Streib et al., 2020; Lu et al., 2014; Polyzotis et al., 2018). Lastly, monitoring of the solution is important to ensure that the results being generated match the intended design and adhere to the governance that was established in the beginning of the process (Cihon et al., 2021; Kavanagh, 2019). RAI has a strong component of monitoring performance in terms of ensuring that the system produced the expected results. As part of the ongoing maintenance of the RAI programs, there are some concepts to be explored that need to be incorporated into an effective RAI program, such as performance drift (Lu et al., 2014), operational bias review (Satell & Abdel-Magied, 2020) and model training and retraining (Adler et al., 2017). This monitoring may be accomplished with a combination of humans (Rahwan, 2017) and other AI to both provide active insight and judgement as well as allowing for the ability to scale the monitoring through the automation (Zetzsche et al., 2020). It is a continual and iterative process to ensure that the AI system, the models and the data are all fair and effective (Arnold & Scheutz, 2018). If the human managers of AI maintain checks and balances through employing continuous champion / challenger exercises with the goal of providing guardrails against an understood benchmark, then the overall trust of the algorithms may increase (Ameen et al., 2021; Ashoori & Weisz, 2019; Rossi, 2019).

2.4.1 RAI principles

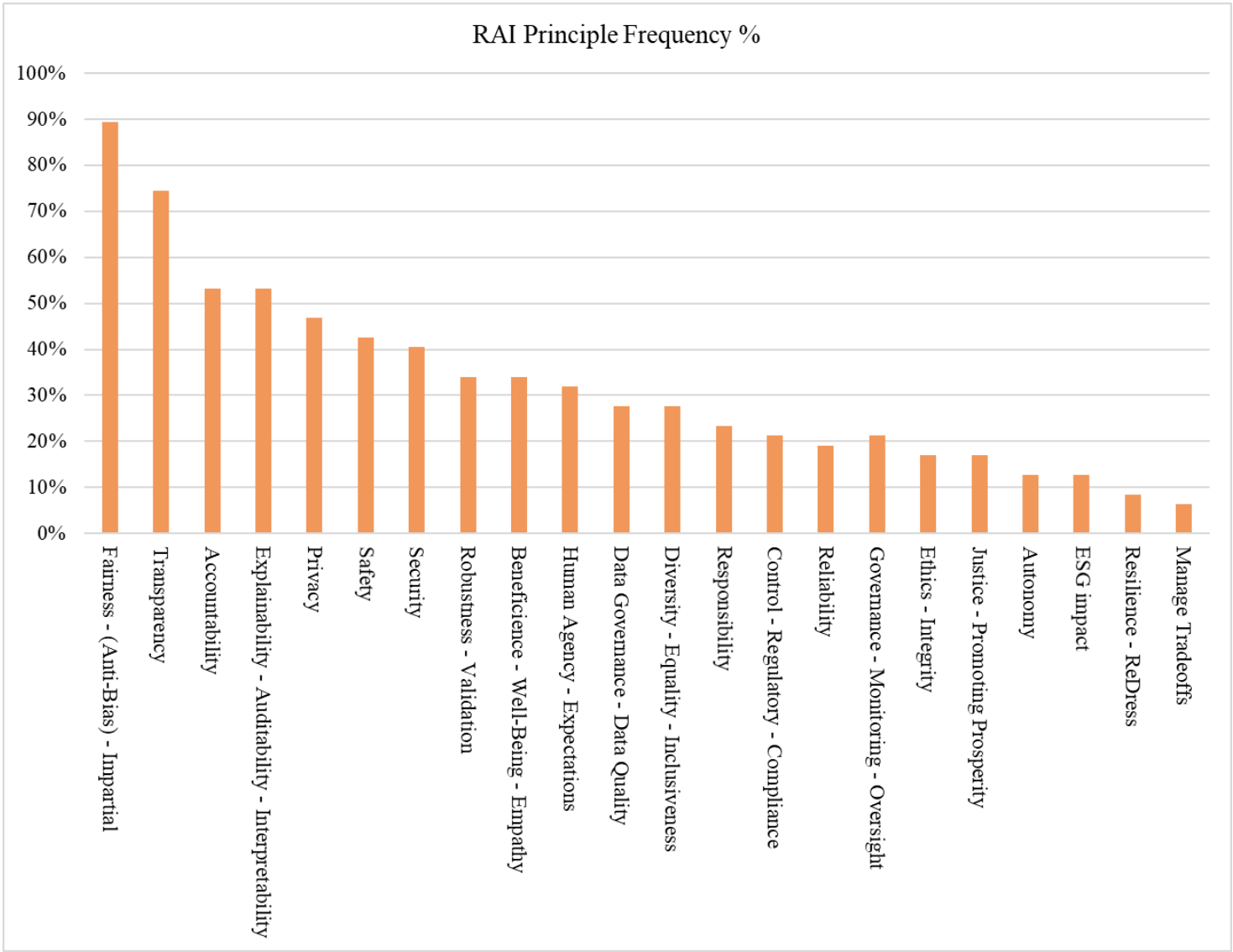
Throughout this research, this study neither identified a standard accepted definition for RAI nor a standard set of principles that comprise RAI. "There was limited consensus even on quite fundamental requirements" (Clarke, 2019). There are many reputable AI engaged companies that have published sets of principles in which this study reviewed and recorded the most common repeated items in **Table 1**.

Many other papers have also created summaries of these principles listed in **Table 1** (Benjamins et al., 2019; Boza & Evgeniou, 2021; Buhmann & Fieseler, 2021; Clarke, 2019; de Laat, 2021; Eitel-Porter, 2020; EUCommission, 2019; Fjeld et al., 2020; Floridi & Cowls, 2019; Floridi et al., 2018; Hagendorff, 2020; Institute, 2022; Jobin et al., 2019; Leslie, 2019; Morley et al., 2019; Myers & Nejkov, 2020; Rakova et al., 2020; Schiff et al., 2020; Siau & Wang, 2020; Zeng et al., 2019). In fact, **Table 1** is similar in nature to a table of principles found in the work by de Laat referencing a framework by the Partnership for AI (Cavello, 2020; de Laat, 2021). This leads to the 'many hands' problem, where responsibility for RAI is distributed and muddled (Schiff et al., 2020). Jobin's compendium, which analyzes and lists 84 sources for RAI principles and guidelines seems to be the most comprehensive and authoritative summary available (Jobin et al., 2019) (**Table 5**), though now dated circa 2019. Another extensive summary by Hagendorff was published in a matrix form similar to the **Table 1** from the primary research (Hagendorff, 2020). Yet, another reference that contains a similar matrix summarizing both principles as well as fairness toolkits was published by the IFC EM Compass (Myers & Nejkov, 2020). **Table 1** is a more concise summary of various relevant papers that have focused on various aspects of the RAI principles and listed their views on the most important elements found pertinent to this study on assessing the maturity of RAI in Banks. The importance of deciding on principles is reinforced in a paper by Benjamins discussing how a corporation should approach applying different principles. (Benjamins, 2020). **Figure 4** illustrates a basic statistical analysis on the principles taken from the named summaries above from Jobin, Hagendorff, and Myers. This study then performed an analysis on the most common elements of the principles as seen in **Table 1** and based the initial hypothesis for the instrument development categories and attributes on the most common items. This resulted in the five categories (organizational commitment, transparency, fairness, data mgmt. and security) that are included in the MRAI survey instrument defined later in this study. Three of the categories (transparency, fairness, security) align directly with that of Jobin, Hagendorff, and Myers, and the collective consensus from the group of relevant papers in **Table 1**, and the other 2 categories (Org commitment, Data Mgmt.) are partially aligned to other elements identified in said studies. The following sections will describe the

instrument development and the rationale for the elements included in the instrument with references to the relevant literature for the importance and applicability to assessing the maturity of RAI.

Figure 4: RAI principles analysis

The diagram below depicts the statistical analysis of key RAI principles from **Table 1** that are referred to in various research papers and demonstrates the frequency of some of principles reviewed.



2.5 Instrument Development Overview

The proposed survey instrument for measuring the MRAI (Maturity of Responsible AI) components as explained above is based on primary research RAI principles as well as researching relevant literature on the related topics. As indicated in **Table 1**, this new MRAI instrument incorporates the pertinent principles and adds another level of granularity to assess the focus on the maturity of the principle. This proposed instrument is similar in nature to one preceding bias questionnaire instrument found in the research literature, which conducted a limited pilot study on general bias governance (Coates & Martin, 2019), and follows a similar development methodology guided by Singh and Smith (Singh, Franceschini, & Smith, 2006) including the pre-validation steps. Another survey was influential as well and was focused on an industry review of software companies working with AI (Vakkuri, Kemell, & Kultanen, 2020).

In order to derive the most relevant attributes or principles to build a survey instrument, this study has researched many firm's RAI principles (**Table 1**), as well as incorporated other summary reviews of principles (Benjamins et al., 2019; Boza & Evgeniou, 2021; Buhmann & Fieseler, 2021; Clarke, 2019; de Laat, 2021; Eitel-Porter, 2020; EUCommission, 2019; Fjeld et al., 2020; Floridi & Cowls, 2019; Floridi et al., 2018; Hagendorff, 2020; Institute, 2022; Jobin et al., 2019; Leslie, 2019; Morley et al., 2019; Myers & Nejkov, 2020; Rakova et al., 2020; Schiff et al., 2020; Siau & Wang, 2020; Zeng et al., 2019) listed in **Table 1**. This research led to a focus on the most relevant repeated principles across the studies, which are incorporated into the instrument survey categories (**Table 1**). In terms of the overall 'governance and accountability' principles that are commonly present in other research, this study has captured, characterized, and expanded these into a new principle called 'Organizational Commitment', as the hypothesis is that those companies with strong executive leadership support of RAI will have the most mature RAI programs (Ransbotham et al., 2019). This category of Organizational Commitment and its attributes is a novel contribution to the RAI principles standard and differentiate the survey instrument from other RAI assessment frameworks. The MRAI instrument is comprised of five main categories, with each containing a few questions within them to elucidate the various attributes of each category.

1) Organizational Commitment
2) Transparency
3) Fairness
4) Data Management
5) Security

The explanation below is similar in format to the work at Harvard by Fjeld, et al (Fjeld et al., 2020), which detailed each of the components for inclusion rationale. Each of these main categories are comprised of sub-attributes that will capture the detailed maturity of the RAI program. There are also a few different examples of survey tools that were found in the research that have set a precedent for a standard or index for assessing the MRAI of a company (Ayling & Chapman, 2021; Boza & Evgeniou, 2021; EUCommission, 2019; Mills & Duranton, 2021). One challenge with setting a single standard for this topic is that there are many types of AI (Boden, 2016) and an overall assessment for RAI may have to be high level and generic. In the case of this study, the focus is on credit lending models and commensurate responsibility of the corporation in terms of fairness in lending. Based on the development of the proposed survey tool instrument that incorporated primary literature review research as well as posing hypotheses, this study will perform a pre-validation of the survey categories and attributes through interviews with the Bank's MRM (Model Risk Managers). The feedback will be incorporated into ensuring that the instrument has robust construct, face, and content validity. As part of the instrument validation, the study will use CFA (confirmatory factor analysis) to ensure that the instrument is valid and reliable. In the following sections, this study reviews the categories as well as the detailed survey attributes to provide the context into the rationale for including these elements in the survey instrument.

2.5.1 Organizational Commitment & Accountability

Understanding how to leverage AI in an organization requires a broader integration of the social environment with which the AI operates (Cihon et al., 2021; Rahwan et al., 2019). To make the most of prediction machines, a company needs to focus on reward functions throughout the organization to better

align with the true goals (Agrawal et al., 2018). Theoretically, the clear organizational commitment components do not have a direct mapping to the core concepts within this section. In terms of building a culture of RAI, there are several attributes that are critical to the success, which are defining a specific organization that bears the name Responsible or Ethical AI, designing the function such that there is significant investment, and measurement of associated decisions and financial return, structuring the group to have meaningful oversight and inclusion from top executives, and providing training for employees. Evolving the culture of the organization to embrace and leverage AI is paramount to developing a strong RAI program (Murphy & Largacha-Martínez, 2021). There are evolving models of how organizations will leverage AI and human talent and manage the collaboration of the two capabilities (Daugherty & Wilson, 2018; Shestakofsky, 2017). At least in the next decade, corporations will leverage AI more for augmentation than replacement of management executives, as emotion, and intuition supplemented are needed in conjunction with amazing AI to address the competitive environments (Ferràs-Hernández, 2017; Jarrahi, 2018). A key goal must be to reimagine the future of work and decision-making to leverage the combined unique capabilities from AI and humans (Jarrahi, 2018). Automation has already impacted 90% of the blue-collar jobs that were performed at the turn of the 20th century, and as we approach the middle of the 21st century, machines are moving into the cognitive white-collar domain (Kelly, 2012). Research has shown that successful AI organizations will shift from siloed work to interdisciplinary collaboration, and the decisions will be more data-driven than experience-based, and lastly that the organization must evolve to embrace an experiment-minded agile and adaptable approach (Fountaine, McCarthy, & Saleh, 2019). In the near term, the key imperative to focus on with AI for organizations is balancing the automation with the augmentation as machines and humans complement each other to create a work environment that is better suited for all stakeholders (Daugherty & Wilson, 2022; Makarius et al., 2020; Raisch & Krakowski, 2021).

It is clear that the models of current work and the roles of humans and machines will continue to evolve quickly as the powerful AI capabilities further permeate the day-to-day business operations

(Hunter, Sheppard, Karlen, & Baliero, 2018). Questions about the balance of roles and tasks today and tomorrow continue to create a significant degree of concern from humans from many angles; and if or when the machines will provide holistic substitution for human work (Balasubramanian et al., 2019). Over the next decade, scholars predict that this mixture will have a gradual impact on humans because as AI capability grows, there will be complementary human roles, such as 'AI auditor' (Coeckelbergh, 2020), that will continue to grow (Agrawal et al., 2019; Shestakofsky, 2017).

Organizations that are committed to RAI may have a dedicated team or COE (center of excellence) that manages the various inputs and outputs related to AI, such as data management, algorithms, models, and data pipeline lifecycles (de Laat, 2021). To have such resources deployed to RAI, companies must commit significant investment, align corporate social responsibility and organizational purpose to this cause, and need to have a positive ROI (return on investment) (Borg, 2021; Clarke, 2019; Minevich, 2020). Corporations that are dedicated to RAI will also make investments in training for their employees in order to ensure that the general concepts are understand and AI is part of the working language of their business practices (Cihon et al., 2021). Lastly, companies that are serious about gaining competitive advantage from investments in RAI will have significant involvement from operating committee leadership and the board of directors with regular briefings and formal scorecards in place (Burkhardt et al., 2019).

2.5.2 Transparency, Explainability & Interpretability

One of the critical aspects within RAI is the transparency with which the decisions that are made leveraging the AI capability can be explained (Bussmann et al., 2020). The lack of algorithmic transparency is one of the main barriers for the wider adoption of AI-based solutions in credit risk management (Misheva et al., 2021). Explainability remains one of the most focused on areas within RAI and is the core of governance within the powerful capability (Cihon et al., 2021). Not everyone, however, believes that algorithmic transparency will solve all explainability issues, in that there are potentially some negative impacts and minimally limitations of transparency as well (Ananny & Crawford, 2016).

Closely related to explainability is interpretability (Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019) of the models, each of which are said to increase algorithmic fairness, transparency, and accountability (Broniatowski, 2021; Burgt, 2019; Cath, 2018). Broniatowski and Miller both suggest a social science lens when analyzing transparency, explainability, and interpretability (Broniatowski, 2021; Miller, 2019). One of the key components of governance is ensuring that there is a clearly defined policy in place for the organization to follow. In this case, there is a legal aspect as well, which can be subject to regulation regarding the disparate impact of the policy (Chen, 2018; Golbin, Rao, Hadjarian, & Krittman, 2020; Hall et al., 2021). As discussed above, the training of employees on this policy and reinforcement of it is part of the RAI corporate culture (Dignum, 2019).

There is broad speculation that there will be formal regulation introduced to mandate transparency in the form of explainability and model auditing processes is what is being coined as Algorithm Audit (Crosman, 2022; Koshiyama et al., 2021; Raji et al., 2020; Wyden, Booker, & Clarke, 2022). In fact, the notion of the term and capability of a 'regulatory sandbox' is gaining traction as a place where models can be demonstrated and approved before they may be released into production mode and especially prevalent in the fintech space where Banks are experiencing fierce competition (Brown & Piroska, 2021; Goo & Heo, 2020). Model auditing through checklists, questionnaires, documentation, model card reporting (Mitchell et al., 2019), and administrative processes have been employed as part of explainability and governance efforts in order to ensure that the intent of the model is indeed what transpires during the execution of the capability (Ayling & Chapman, 2021). There have been a number of auditing tools introduced into the AI/ML processes as general awareness of capability and risk of scaling AI models has risen (Bussmann et al., 2020). One of the tools that is gaining popularity with practitioners that are focused on explainability is the use of knowledge graphs to provide a visual representation (e.g. visualizing hidden states of a neural network) of the model, which assists with trust, accuracy and productivity in providing the model explanations (Tiddi & Schlobach, 2022). Another concept that is important in the context of explainability is the notion of model drift, in which various environmental factors may render some part of the ML lifecycle obsolete (Barros & Santos, 2019;

Demšar & Bosnić, 2018; Lu et al., 2014). In cases where model drift has been identified, model reparation must be performed to remedy whichever component has the situationally, identified resident flaw (Davis, Williams, & Yang, 2021) .

A number of the use cases above pertain to supervised (models that have data labels) ML models, which are the more understood, and frequently used models in Banks due to existing regulation; however, there is an advanced ML capability with unsupervised ML models in which the models are effectively 'black boxes', and are difficult to audit, requiring more advanced explanatory techniques to provide the interpretability of the models (Adler et al., 2017). In the case of more advanced ML, equally advanced explanatory technique is required in the form of emerging tools such as LIME (Ribeiro, Singh, & Guestrin, 2016) and SHAP (Lundberg & Lee, 2017). These advanced explanation tools provide additional visibility into the opaque nature of the unsupervised ML decision-making output (Gramegna & Giudici, 2021). With the detailed elements discussed below, this study incorporates these items into the MRAI instrument to assess the maturity of RAI transparency, primarily composed of the ability to demonstrate explainability and interpretability in the data, models, and ML algorithms.

2.5.3 Fairness and Bias Mitigation

Bias is inherent in human thinking and an unavoidable characteristic of data collected from human processes (Dignum, 2019). The human decisions, however, are not only biased but also noisy and inconsistent (Cowgill, 2019). At the core of this research is assessing the maturity of RAI practices, which focus on the mitigation of bias in credit lending decisions. Some of the key attributes associated with measuring the degree of fairness are policy review, mitigating proxy discrimination, assessing training data management, understanding to what extent humans are involved in the decision-making process, legal and regulatory considerations, and action plans for when the models and algorithms run amok. Bias manifests itself in scenarios where the result or action of a decision is perceived as unfair to specific groups or individuals (Arnold & Scheutz, 2018). There are many examples in society and life where bias exists and is captured and stored in the form of data. Criminal records, bill payment history,

education history, and address location are prime examples (Coeckelbergh, 2020; Kearns & Roth, 2019; O'Neil, 2016). These data are then processed, trained, run through AI/ML and correlated against many variables to assess whether one is accepted for whichever application they may have applied, and when the data are incorrectly leveraged, resulting in bias (Anderson & Anderson, 2007). A key element of AI/ML is that in order to make the predictions, the technology applies statistical analysis to large volumes of data, which can be devoid of the richness of sentience (i.e., lacking the necessary contextual knowledge) (Balasubramanian et al., 2019). This architecture which is designed by humans and leverages data representing attributes of human profiles bears inherent risks for bias, privacy, and misuse (O'Neil, 2016; Teichmann, 2019).

To provide some additional context, this study refers to some other examples, the first of which is related to recidivism and the discrimination resident in the COMPAS data (Fry, 2018). There are examples in many categories that describe biases, such as banks making credit decisions (Roszbach, 2003) about certain individuals residing in certain zip codes (Žliobaitė, 2017), passing over candidates for hiring due to undesirable data profile attributes (Tambe et al., 2019), and customer service virtual agents mishandling requests due to incorrect context from the AI (Hunt, 2016). These examples carry risk and require RAI governance to mitigate implications (Coeckelbergh, 2020).

In terms of assessing the existence of specific fairness policies, it is a review of guidelines, training, and executive reinforcement to engrain responsible and fair AI into company policies, guidelines, processes and models (Mitchell, Potash, Barocas, D'Amour, & Lum, 2020; Rao & Golbin, 2019). The importance of policies is brought to life by Renda and the CEPS task force in their AI governance paper (Renda, 2019). Provisions within GDPR already provide the right for individuals to understand the inner workings of the ML models and obtain explanations for the ML outcomes (Alves, Amblard, Bernier, Couceiro, & Napoli, 2021). There is also a legal aspect as many discrimination laws are already in place for pre-ML credit risk practices; however, with the advent of scaled AI and ML take a more threatening shape, and Koshiyama suggests there will be more formal algorithm audits (Koshiyama

et al., 2021; Wyden et al., 2022). A unique technique for performing algorithmic audits takes a page from security practices where a bounty is offered for finding security vulnerabilities. In this case, if one can find a bias issue, then a bounty is awarded, as in the Twitter algorithmic bias bounty challenge (Chowdhury & Williams, 2021). An important distinction in ML is that it is a prediction technology, and can either have an automated decision-making element or leave the judgement and decision to a human (Agrawal et al., 2018). In terms of building an instrument to survey MRAI capability, the research suggests it is important to include the human element in what is known as 'Human In The Loop', which allows for a human to infuse context and judgement into decisions (Zetzsche et al., 2020). There are also cases where the bias and discrimination is not specifically intended, and rather found incidentally through bias in training data or other data elements, which can lead to proxy discrimination (Prince & Schwarcz, 2020). In order to try to prevent bias, a relatively new technique is ensuring that ML models are accompanied by 'model cards', which provide some explanatory documentation, that detail the expected performance characteristics of the model (Mitchell et al., 2019). Bias and discrimination can also arise as a result of model drift, which again is not intentional, but rather a function of environmental, data, or context changes occurring, which create scenarios of intended consequences (Barros & Santos, 2019; Demšar & Bosnić, 2018; Lu et al., 2014). Nevertheless, if these scenarios do occur, it is important to have the explainability and transparency to be able to detect and remediate the bias through model reparation (Davis et al., 2021).

2.5.3.1 Regulation

As a result of this recognized potential for bias and discrimination (Prince & Schwarcz, 2020), the regulatory agencies have enhanced the perspective on the Fair Credit Reporting Act (FCRA), enacted in 1970, and the Equal Credit Opportunity Act (ECOA) (ftc.gov, 2020a, 2020b), enacted in 1974, added more regulation around hiring practices that leverage AI for speed and efficiency of screening candidates (Friedman & McCarthy, 2020; Maurer, 2020). Regulation has been applied to these lending decisions since the 1970's (Candelon et al., 2021). Agencies in a couple of categories related to discrimination, such as the Fair Credit Reporting Act (FCRA) (Fay, 2021), FHA (Fair Housing Act) (Lee & Floridi,

2020) and the Equal Credit Opportunity Act (ECOA) (ftc.gov, 2020a), have been created to assist with governance. More recently, there was a proposal to Congress, named the Algorithmic Accountability Act of 2019, which did not pass, but set in motion what some deem as inevitable (Burt, 2021; Congress.gov, 2019). In fact, there is a new push for the Algorithmic Accountability Act in 2022 (Wyden et al., 2022). The importance of this topic is recognized in a research paper by Almeida that summarizes a number of papers that discuss AIR (AI Regulation) in the prior decade (Almeida, dos Santos, & Farias, 2021).

Regulation also supervises the risk management of the Banks as well as the risk of harm to the individuals in a constantly monitored capacity (Barth, Lin, Ma, Seade, & Song, 2013). SR (Supervision and Regulation) 11-7 (Supervisory Guidance on Model Risk Management) from the Fed and OCC (Office of Comptroller of the Currency) explicitly defines the risk management around credit lending (FederalReserve.gov, 2011a, 2011b). It focuses on ensuring Banks have model validation processes (Lorica, Doddi, & Talby, 2019a) instituted to be able to demonstrate conceptual soundness, monitoring, and outcome analysis (FederalReserve.gov, 2011b). There is also regulation that is focused on protecting the consumer and avoiding harm that is managed by the CFPB (Consumer Financial Protection Bureau) and UDAAP (Unfair and Deceptive Act Practices) (CFPB, 2011). Since AI/ML is now leveraged ubiquitously in assisting with these credit underwriting and lending decisions (Roszbach, 2003), the regulatory scrutiny that applied to the AI processes is trending toward governance by RAI practices. With the significant advancements and rapid acceleration in the technical capability of AI, humans have realized that instituting governance, certification and code of conduct programs to regulate (Etzioni & Etzioni, 2017) ethical and fair use of AI should be now mandated (ftc.gov, 2021; Zetzsche et al., 2020).

The power of AI is truly transformational in terms of automating tasks that are performed inefficiently by humans and scaling the ability to process those tasks by orders of magnitude (Raisch & Krakowski, 2021; Shestakofsky, 2017). This power in trusting AI is then magnified when employing ML techniques due to the lack of intervention from humans in enhancing the AI models (Jordan & Mitchell, 2015). Consequently, many companies and the various interacting stakeholders have realized the double-

edged sword of this amazingly powerful technology and begun to establish AI governance programs to monitor, manage and remove bias and error in the programming treatments (ftc.gov, 2020b; Rakova et al., 2020). These governance programs are part of a movement around ethical and RAI, and are being mandated by customers, employees, as well as governmental regulatory authorities (Anderson & Anderson, 2007; Leslie, 2019). An important point in the development of the governance programs is that due to the complexity of many components in the overall process of AI, the guidelines need to ensure and monitor the chain of responsibility across all of the environmental actors (Dignum, 2019).

2.5.4 Data Management

In order to effectively scale AI, large volumes of data must be processed with immense computing power (Jordan & Mitchell, 2015). It is critical for AI to have a statistically valid data set with enough volume that it represents a meaningful amount of the group that is being studied (Polyzotis et al., 2018). AI systems use data that we generate in our daily lives, mirroring our varied attributes which make it susceptible to containing bias (Dignum, 2019). It must be acknowledged that the availability, quality, and structure of the data are the key attributes for the foundation of AI (Duan et al., 2019). The theoretical conditions necessary to completely eliminate bias are extreme and unlikely to appear in real datasets (Cowgill, 2019). One challenge that belies big data (Mayer-Schnberger, 2013) and AI models is that they currently rely on historical data to train the models, and thus there doesn't seem to be foresight about what may evolve in the future, and the correct answer for the AI model is based only the past data (O'Neil, 2016). Banks are challenged to organize the internal data that they manage and the data are usually fragmented across different collection processes, systems and organizations throughout the bank (Tammenga, 2020).

Data training in AI is perhaps one of the most important concepts to ensure that bias is not present within the overall AI data lifecycle as even a pristine model performing exactly as designed can create unintended consequences if the data that the model uses to train is biased (Hall et al., 2021). There are data reparation techniques that can be applied to correct the data (Davis et al., 2021) in cases where there

are identified inequalities that permeate the AI lifecycle. The notion of synthetic data also exists where the corporation will infuse manufactured or fabricated data into the environment to influence the models, which is used when there is not sufficient nor clean training data on which to base the model training (Campbell, 2019; Gupta, Bhatt, & Pandey, 2021; Vanian, 2021). Coinciding with the concept of synthetic data is the GAN (generative adversarial network) capability, which is used in cases where there are insufficient training data (Creswell et al., 2018), thus creating an intentional tension within the program between the generator and the discriminator to ensure that the training data will serve the intended purpose when the model is productionalized.

AI has also experienced an evolution of its own in maturing the capability from supervised learning (trained with specific structured labeled data sets) to unsupervised learning (trained with ML on expansive unstructured unlabeled data sets) (Jordan & Mitchell, 2015; Loukides, 2016; Zador, 2019). Algorithms are backward looking because they have been trained on data from the past (Polyzotis et al., 2018). To provide for evolving models that keep pace with societal evolution, continuous maintenance and human intervention will be needed (Hunter, Sheppard, Karlen, & Balieiro, 2018; Rahwan, 2017; Zetzsche et al., 2020).

Another important concept in the data and models is explainability or model transparency which is simply the ability to explain the results of an analysis provided by a model based on some data (Holzinger et al., 2018; Stoyanovich et al., 2020; Whang, Tae, Roh, & Heo, 2020). One of the risks with advanced AI models using unsupervised and reinforcement learning is that they may be black boxes (Adler et al., 2017) in which human operators do not fully understand how the AI derived the analytics that were produced. There are even extreme situations where the data labels are manipulated, which is known as data poisoning (Jagielski et al., 2018), and this has significantly impacts the effective performance of AI. In the cases of prediction of credit loan defaults (**Figure 2**), or recommended products for cross-selling, there is a belief that humans should be able to understand the rationale for the AI decisions (Anderson & Anderson, 2007; Kearns & Roth, 2019). Privacy and Regulation with the EU's GDPR (General Data Protection Regulation and GINA (Genetic Information Nondiscrimination Act) also

are important items to consider when instituting AI due to the nature of the data being utilized (Tambe et al., 2019; Wachter et al., 2018). Of specific note, there is interest in differential data privacy in which data structures and patterns are disclosed, however, certain PII (Personally Identifiable Information) data elements are withheld or encrypted (Al-Rubaie & Chang, 2019; Dwork, Rothblum, & Vadhan, 2010). In addition to data privacy, there is the notion of the 'right to be forgotten', in which one would like their data deleted from the systems (Tjong Tjin Tai, 2016). There are several considerations and dependencies that AI has on the data, and thus the management of clean, minimally biased, quality data is paramount to have effective AI (Polyzotis et al., 2018).

Data or concept drift can also occur in the ML models, as the data naturally change in the environment, the original intent of the model may no longer be valid and reparation must be performed to re-align the model for effectiveness (Davis et al., 2021; Lu et al., 2014). More mature data management operations that practice RAI have a few tools in place known as DataOps, which is based on DevOps for programming, but applied to data management (Rodriguez et al., 2020). In addition to practicing DataOps, mature data management operations will have established data pipeline management tools to manage the full lifecycle of ML, with the cleansing, pre-processing, alignment, training, usage, and post-processing of the data (Deepa & Ramesh, 2021). Another tool that is leveraged for advanced data management is an EDA (exploratory data analysis) toolset, which provided a visual capability for analyzing the data (Hafen & Critchlow, 2013).

Another critical component of AI and ML is ensuring that the data that are being used throughout the pipeline lifecycle remain private and secure (Papernot et al., 2016; Wolf, 2020). One of the methods for ensuring data privacy and security is to apply encryption (Rist, 2018) and access control to the databases and data (Wee & Nayak, 2019). Along with the organizational impact and commitment referred to above, it is critical to have a CDO (Chief Data Officer) involved with the details of the data management and reviewing components of the process to ensure the data are appropriately prepared and maintained for the AI/ML processes (Dell, 2020).

2.5.5 Security

Elements of security to prevent intrusion and guard vulnerabilities are paramount to ensuring fairness, safety and privacy in RAI (Papernot et al., 2016). In terms of security elements, this paper views as important components of MRAI, the focus is on capabilities that mitigate intrusions, data encryption, processes for handling models should they become infected or fall into unintended possession, ability to control the algorithms, and special security for production runtime environments. This study touched on the data privacy topic (including differential privacy) in the data management section, but one of the key data elements to protect with security is PII (personally identifiable information), such as name, credit score, social security number and related sensitive personal information. Perhaps the largest concern in terms of security is cyber adversarial attacks named AML (adversarial ML) where malevolent actors are breaching system or network security to gain illegal access to various protected assets (Anthi, Williams, Rhode, Burnap, & Wedgbury, 2021; Wee & Nayak, 2019).

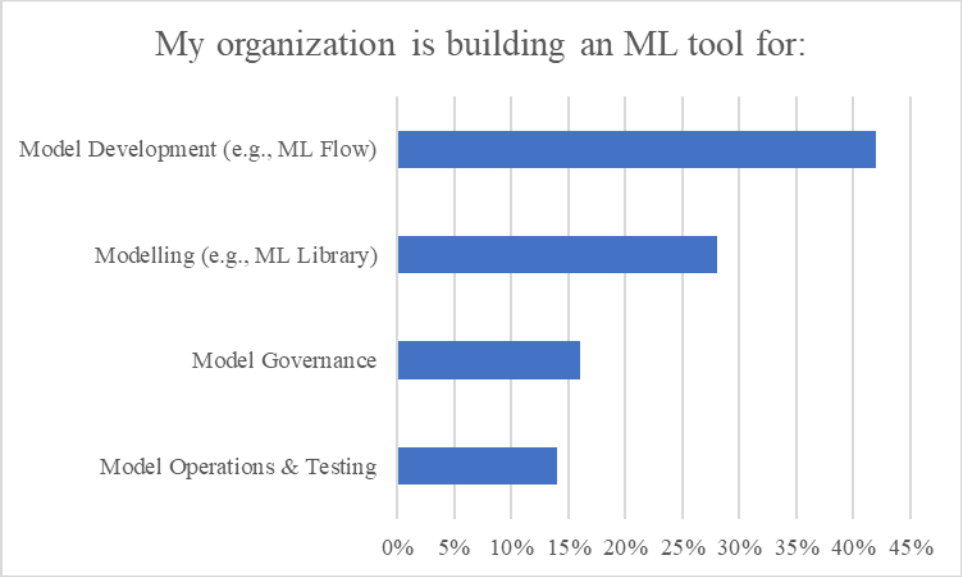
Securing the data is important for multiple reasons. First, there is the notion of data poisoning, where a malicious actor will break the training data or production data and insert some data, which results in a data poisoning attack (Jagielski et al., 2018; Wang et al., 2019). This is important because as described in the prior sections, the ML model will integrate and learn what it ingests from the data, so if inappropriate data is introduced into the data lifecycle, varied unpredictable results could ensue. On a related note, additional support for good explainability relates to security, as if models are fully explainable, and results are expected, then if there is a particular breach and unexpected outcomes begin to occur, a Bank can assume that they minimally need to determine if there has been a breach. If there is little to no explainability in terms of the expected outcomes of the models, then a breach could go undetected (Barredo Arrieta et al., 2020). In the case of a breach, there are also techniques actually leveraging ML to detect the database intrusion and suppress the actor (Wee & Nayak, 2019). Testing and verification for security is a key indicator of maturity in the AI/ML data and security environment, and we are interested in the verification procedures that Banks have in place for this capability (Papernot & Brain, 2018).

2.6 RAI Assessment Frameworks and Toolkits

In addition to the various RAI principles that have been published, there are also a number of assessment frameworks and toolkits that have been developed in an attempt to provide tools to audit models and prevent bias (Aequitas, 2019; AIEthicist, 2021; Bellamy et al., 2018; Deon, 2021; GooglePAIR, 2021; IBMAI, 2021; MetaAI, 2021; MicrosoftFairlearn, 2021; Saleiro et al., 2019). A comprehensive list of principles, assessment frameworks, and toolkits is listed on the AI Ethicist website (AIEthicist, 2021) as well as another summary listed on the Medium website (Durmus, 2021). This study has included a few of the popular toolkits recently available from some of the larger technology firms in **Table 6**. The toolkits range from simple questionnaires and checklists to assessment tools to full data ingestion code-based (e.g., Python, R) software on GitHub (GitHub, 2022) that generate bias interpretation based on data ingested into the tool.

Some of the questions present in the toolkits and assessments are relevant for this study's methodology in terms of prioritized attributes to include in the survey instrument. The proposed instrument survey questions all refer to the key RAI principles from the prior section. This paper discussed regulation in an earlier section and has included a question around regulatory sandboxes in terms of assessing the maturity of the RAI. Some of the assessment toolkits are in fact API (application programmer interfaces) that can be leveraged to validate the models in a regulatory sandbox environment. These tools are software code that one conversant in the ML tooling (**Figure 5**) could operate in a regulatory sandbox test environment. An interesting development is that a number of these assessment tools are being developed, which adds to an array of validation and audit tools that can be leveraged to assess the fairness of ML models (Durmus, 2021). One significant advance in tools and maturity of RAI in this genre is the recent introduction of a certification program for RAI that was developed by the RAI institute (Institute, 2022).

Figure 5: ML Tool usages



Source: O'Reilly Survey - https://www.oreilly.com (Lorica, Doddi, & Talby, 2019b)

2.6.1 Capability Maturity Models

Leveraging the well-known framework of CMM (Capability Maturity Model), this study has designed a maturity assessment instrument for the RAI capability maturity. As an output of performing the survey administration and collecting responses of the varied elements for each category, this study is effectively able to derive a capability maturity model (Paulk, Curtis, Chrissis, & Weber, 1993) with the data from the survey. There has been discussion of a maturity model in some other research works in terms of calling for an RAI maturity model to be developed (Vakkuri et al., 2021) as well as literature reviews summarizing the research landscape (Sadiq, Safie, Abd Rahman, & Goudarzi, 2021). There are also companies that are offering a capability to perform and document the assessment (Fifth-Quadrant, 2021). In the case of the Fifth-Quadrant RAI index, the maturity model contains 4 categories (Planning, Initiating, Developing, Maturing), which deviates from the classic 5 tier CMM (Initial, Repeatable, Defined, Managed, Optimizing) referenced above. Another example from Element AI contains the 5 categories, however, the authors have changed the names slightly to Exploring, Experimenting, Formalizing, Optimizing, Transforming (Element.AI, 2021). In the case of this study, the instrument

Likert scale contains 5 potential assessment levels building on the classic CMM archetype and will be depicted with the collected study data with categories from **Figure 6**. A last example from Gartner (Gartner & Panetta, 2019) that created a CMM model for AI with the following categories (Awareness, Active, Operational, Systemic, Transformational) is the most appropriate model for this study to employ.

Figure 6: Gartner AI Maturity Model

This study has adopted the Gartner categories (Gartner & Panetta, 2019) for the **MRAI CMM** model.

AI Maturity Model				
Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Awareness	**Active**	**Operational**	**Systemic**	**Transformational**
				AI is part of the business DNA
			AI is pervasively used for digital process and chain transformation, and disruptive new digital business models	
		AI in production, creating value by e.g., process optimization or product/service innovations		
	AI experimentation mostly in a data science context			
Early AI interest with risk of overhyping				

Source: Gartner 2019

2.7 Instrument Development Summary

As has been explained above, the attribute elements that comprise each of the 5 categories provide a level of granularity in the survey for which if a Bank has high scores in each of these subitems, they will derive a high score for the category and demonstrate the degree of maturity they possess within that particular component of RAI. With the survey instrument categories and detailed attributes being defined, this study endeavors to review a pre-validation survey instrument with the survey population to obtain pre-validation feedback in terms of face and content validity on the robustness and comprehensiveness of the survey questions. Upon receiving feedback from the initial set of interviews, the feedback will be incorporated into the final post-validation survey instrument to be used with the broader population of Banks described in the Data section. As mentioned above, since this study is focused on ML for credit lending, the RAI survey categories chosen are a combination of elements that could pertain to any company deploying RAI as well as certain elements that are very particular to data, algorithms, and models that are leveraged in credit underwriting and lending decisions. In addition to this survey, as a supplement, there will also be primary research performed directly on a set of publicly available information to develop a Proxy MRAI score, which will be described in the next section.

2.8 Proxy MRAI Score

In order to provide another level of validation to the instrument, this study will also research several Bank attributes that are publicly available to develop a proxy MRAI score. It is anticipated that the large Banks will have plenty of data to attest to the various attributes of the proxy MRAI score, however, potentially some of the smaller banks may not have this data publicly available. In this case, we assume that if there is no public evidence for the attribute being true or false, that it is false. The proxy MRAI score will be derived from publicly available information on web sites and corporate documents (annual report, 10k, shareholder letters).

The questions for the attributes being recording in the proxy MRAI score are:

1) Does an in-house research team exist for R(AI) and/or are there research publications available?

2) Are there R(AI) articles in the press published from the company?
3) Are there a set of published R(AI) principles available for the company?
4) Is R(AI) mentioned in the corporate documents (e.g., – annual report)
5) Is there a direct link from the website to the R(AI) perspective or philosophy?
6) Are there published academic research partnerships for the company?
7) Does a R(AI) or related capability COE (center of excellence) exist for the company?
8) Are there published careers within the R(AI) field available for the company?

Each of these will be scored in a binary format and then recorded for overall strength of a score of 0-8 depending on the elements that are publicly available.

2.9 Survey Deployment

There will be 3 components of the survey for 50 of the top US Banks. First the pre-validation survey was developed based on the RAI principles and then interviews performed with the Banks. This survey is focused on reviewing face and content validity of the categories as well as obtaining feedback on the actual categories. To provide a robust analysis, this study will also perform Cronbach's alpha and CFA analysis on the items of the pre-validation survey. Second is the validated MRAI survey instrument, which will be administered to the Model Risk Managers, and thirdly is a set of ESG (Environmental, Social, Governance) questions, which will be captured to provide additional data for an MTMM (Multi-Trait Multi-Method) construct validity analysis.

Based on finalizing the pre-validation survey and in addition to conducting the proxy MRAI scoring study, this study will administer the instrument in survey interviews to the set of 50 Banks in the research scope that will provide additional insights on the maturity or the RAI program from a Bank self-assessment capacity. This interview survey will be conducted with an executive in the Model Risk Management area of the Bank and typically an MRM (Model Risk Manager) which will record a self-assessment of the company's status in a [5-point Likert] format (**Table 2**) for the various elements of the survey. At the end of the interview, the study will also ask a few questions about the ESG strategy and

leadership perspective. This analysis will produce three sets of Cronbach's alpha and CFA internal consistency reliability statistical analyses. The first set of statistics will measure the reliability of the pre-validated instrument design categories. The second set of statistics will measure the collection of the Bank data with the post-validation instrument. The third set of statistics will measure the Bank's ESG strategy for inputs into the MTMM analysis. With these statistical analyses performed, this study will produce a robust and valid MRAI instrument.

As described in the literature review, RAI is a relevant topic for today's growing use of AI. A gap exists in the consistency both in definition as well as application of RAI principles in the literature today. In addition, the ability to validly measure a Bank's RAI does not really exist, or if it does is not public and likely in obscure methods not recognized by the regulators. This study described the components (categories and attributes) in comprehensive detail in the above section providing rationale for inclusion in the tool and referenced relevance from extant literature. The advent of this MRAI instrument will provide Bank's a mechanism and toolset with which to measure the maturity of their RAI program and capabilities.

Table 1: Summary of Research papers on principles

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
Accenture	Eitel-Porter, R. (2020). Beyond the promise: implementing ethical AI. AI and Ethics, 1(1), 73-80. https://doi.org/10.1007/s43681-020-00011-6	X	X	X	X	X																	
AI4People	Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V.(2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach (Dordr), 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5			X	X					X									X	X			
ALTAI (EU – European Commission High-Level Expert Group)	The Assessment List for Trustworthy Artificial Intelligence https://altai.insight-centre.org/	X	X			X	X		X	X	X						X				X		
Amazon	Fairness, accountability, transparency, ethics https://www.amazon.science/tag/fairness-accountability-transparency-ethics-fate	X	X	X														X					
Asilomar principles	Future of Life Institute http://futureoflife.org/ai-principles	X	X			X					X			X	X				X				
Bain	Data Scientists, Take a Hippocratic Oath While the ethics of analytical tools can be tricky to parse, five basic principles can help data scientists address the challenge. https://www.bain.com/insights/data-scientists-take-a-hippocratic-oath-forbes/	X	X			X			X	X		X				X							
BCG	Are You Overestimating Your Responsible AI Maturity? https://www.bcg.com/publications/2021/the-four-stages-of-responsible-ai-maturity	X	X	X	X	X	X	X	X			X	X			X			X		X		

Table 1: Summary of Research papers on principles, cont.

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
Business Roundtable	https://s3.amazonaws.com/brt.org/Business_Roundtable_Artificial_Intelligence_Roadmap_Jan2022_1.pdf	X	X		X			X				X	X				X						
Buhmann, Alexander Fieseler, Christian	Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. Technology in Society, 64. https://doi.org/10.1016/j.techsoc.2020.101475	X					X	X	X	X			X					X	X	X	X		X
Cap Gemini	Ethical AI – Decoded in 7 Principles https://www.capgemini.com/2021/04/ethical-ai-decoded-in-7-principles/	X	X	X	X	X	X		X			X			X							X	
CitiBank	ARTIFICIAL INTELLIGENCE - AN ETHICAL STANCE https://www.citi.com/mss/solutions/pfss/solutions/fund/fiduciary-services/assets/docs/complexity/innovation/Ethics-of-AI-graphic.pdf	X	X		X			X			X	X					X						
Clarke, Roger	Clarke, R. (2019). Principles and business processes for responsible AI. Computer Law & Security Review, 35(4), 410-422. https://doi.org/10.1016/j.clsr.2019.04.007		X		X		X		X	X		X			X								
de Laat, Paul	de Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Philos Technol, 1-59. doi:10.1007/s13347-021-00474-3	X	X	X	X	X	X	X						X										
Deloitte	Deloitte's Trustworthy AI™ framework https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html	X	X	X	X	X	X	X	X					X		X								
Department of Defense	Defense Innovation Unit Publishes 'Responsible AI Guidelines' https://www.defense.gov/News/News-Stories/Article/Article/2847598/defense-innovation-unit-publishes-responsible-ai-guidelines/	X	X	X						X	X													

Table 1: Summary of Research papers on principles, cont.

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
EY	How we can ensure that AI benefits everyone https://www.ey.com/en_gl/wef/how-to-embrace-ai-responsibly-and-make-it-inclusive	X			X											X		X					
Facebook	Facebook's five pillars of Responsible A https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/	X	X	X		X	X	X	X				X		X		X						
Forbes	Six Essential Elements Of A Responsible AI Model https://www.forbes.com/sites/forbestechcouncil/2021/09/01/six-essential-elements-of-a-responsible-ai-model/?sh=27d726e156cf	X	X	X				X									X					X	
Google	Artificial Intelligence at Google: Our Principles https://ai.google/principles/	X		X		X	X			X								X	X				
Hagendorff, Thilo	Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8	X	X	X	X	X		X					X	X					X	X	X		
Harvard (Fjeld, et al.)	Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420	X	X	X	X	X	X	X						X	X								X
(IEAIML) Institute for Ethical AI & Machine Learning	The Responsible Machine Learning Principles https://ethical.institute/principles.html	X			X	X			X			X								X			

Table 1: Summary of Research papers on principles, cont.

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
IBM	AI Ethics https://www.ibm.com/artificial-intelligence/ethics	X	X		X	X			X														
IEEE	The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems https://doi.org/10.100 7/978-3-030-12524-0_2	X	X	X						X	X								X				
IFC (Myers, Gordon Nejkov, Kiril)	Developing Artificial Intelligence Sustainably: Toward a Practical Code of Conduct for Disruptive Technologies https://www.ifc.org/wps/wcm/connect/publications_ext_content/ifc_external_publication_site/publications_listing_page/emcompass-note-80-tocc	X	X	X				X	X	X			X	X			X					X	
IICP	IICP - The Conference toward AI Network Society https://www.soumu.go.jp/main_content/000507517.pdf		X			X	X				X				X			X					
Informatica	6 Key Principles for Responsible AI https://www.informatica.com/blogs/6-key-principles-for-responsible-ai.html	X	X	X	X							X		X				X					
InfoSys	https://www.infosysconsultinginsights.com/wp-content/uploads/2021/06/implementing-responsible-ai_infosys-consulting_pov.pdf	X		X	X	X				X			X						X				
INSEAD (Boza & Evgeniou)	Boza, Pal and Evgeniou, Theodoros, Implementing Ai Principles: Frameworks, Processes, and Tools (February 10, 2021). INSEAD Working Paper No. 2021/04/DSC/TOM, Available at SSRN: https://ssrn.com/abstract=3783124 or http://dx.doi.org/10.2139/ssrn.3783124	This paper referred to other principles																					

Table 1: Summary of Research papers on principles, cont.

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
Jobin, Anna	The global landscape of AI ethics guidelines https://doi.org/10.1038/s42256-019-0088-2																						
KPMG	Ethical AI: Five guiding pillars https://advisory.kpmg.us/articles/2019/ethical-ai.html	X	X					X									X	X					
McKinsey	Leading your organization to responsible AI https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/leading-your-organization-to-responsible-ai	X			X							X			X		X						
Microsoft	Microsoft AI principles https://www.microsoft.com/en-us/ai/responsible-ai	X	X	X		X	X	X					X			X							
Mittelstadt, Brent	Principles alone cannot guarantee ethical AI https://doi.org/10.1038/s42256-019-0114-4	This paper referred to other principles																					
Medium	Introduction to the 4 Principles of the Responsible AI for Business Leaders https://medium.com/codex/introduction-to-the-4-principles-of-the-responsible-ai-for-business-leaders-b7f5c8df5ba9	X	X						X	X													
Morley, et al.	From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices https://arxiv.org/abs/1905.06876	X	X		X	X	X			X	X	X				X	X			X		X	X
OECD	Recommendation of the Council on Artificial Intelligence https://doi.org/10.1017/ilm.2020.5	X	X	X	X		X	X	X	X	X		X										

Table 1: Summary of Research papers on principles, cont.

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
Partnership for AI	PAI Launches Interactive Project To Put Ethical AI Principles into Practice https://partnershiponai.org/pai-launches-interactive-project-to-put-ethical-ai-principles-into-practice/	X	X	X			X				X												
PWC	Ethical AI: 10 principles the world (mostly) agrees on — and what to do about them https://www.pwc.com/us/en/tech-effect/ai-analytics/how-to-make-ai-ethical.html	X		X	X	X	X	X	X	X	X				X	X							
Responsible AI Institute	Working Together for AI We Can Trust https://www.responsible.ai/	X	X	X	X		X	X	X		X												
Salesforce	AI for Good: Principles I Believe In https://www.salesforce.org/blog/ai-good-principles-believe/	X	X				X			X	X			X									
Siau, Keng Wang, Weiyu	Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. Journal of Database Management, 31(2), 74-87. https://doi.org/10.4018/jdm.2020040105	This paper referred to other principles																					
Stanford	Stanford Institute for Human-Centered Artificial Intelligence https://hai.stanford.edu/	X								X	X		X										
Telefonica	Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by Design in Practice. AAAI Proceedings - Telefonica	X	X		X	X		X			X				X								
Turing Institute (Leslie)	Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529	X	X	X			X	X				X				X					X		

Table 1: Summary of Research papers on principles, cont.

Name	Reference	Fairness - (Anti-Bias) - Impartial	Transparency	Accountability	Explainability - Auditability - Interpretability	Privacy	Safety	Security	Robustness - Validation	Beneficience - Well-Being - Empathy	Human Agency - Expectations	Data Governance - Data Quality	Diversity - Equality - Inclusiveness	Responsibility	Control - Regulatory - Compliance	Reliability	Governance - Monitoring - Oversight	Ethics - Integrity	Justice - Promoting Prosperity	Autonomy	ESG impact	Resilience - ReDress	Manage Tradeoffs
Twitter	Introducing our Responsible Machine Learning Initiative https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative	X	X								X		X	X						X			
World Economic Forum	How to Prevent Discriminatory Outcomes in Machine Learning https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning	X			X								X									X	
Zeng, et al.	Zeng, Y., Lu, E., & Huanfu, C. (2019). Linking Artificial Intelligence Principles. AAAI Proceedings on Artificial Intelligence Safety. https://arxiv.org/abs/1812.04814	X	X	X	X	X	X	X	X			X	X	X	X	X	X						
		41	**34**	**24**	**24**	**21**	**20**	**19**	**16**	**16**	**15**	**12**	**13**	**10**	**10**	**9**	**10**	**7**	**8**	**6**	**6**	**4**	**3**
		87%	72%	51%	51%	45%	43%	40%	34%	34%	32%	26%	28%	21%	21%	19%	21%	15%	17%	13%	13%	9%	6%

Table 2: Survey Instrument

#	Factor Name	#	Key Evidence	Scoring				
1	**Organizational Commitment to RAI**		**This attribute measures the degree of organizational commitment to RAI in terms of org structure, financials, accountability.**					
	Org Structure	1	To what degree is there a formal org structure entity called Responsible or Ethical AI?	1	2	3	4	5
	Investment in RAI	2	To what degree is there evidence of significance financial investment linked to Responsible AI.	1	2	3	4	5
	ROI analysis on RAI	3	To what degree is there a formal financial ROI analysis performed on Responsible AI?	1	2	3	4	5
	Training for RAI	4	To what degree are there training programs in place for all employees on Responsible AI?	1	2	3	4	5
	Culture of AI	5	To what degree is there a perception of a culture of AI within the company?	1	2	3	4	5
	Pareto Efficiency Frontier Decisions	6	To what degree is there evidence of using pareto efficiency frontier analysis in credit lending?	1	2	3	4	5
	C-Suite Involvement	7	To what degree is the CEO or Board updated on the company's RAI program?	1	2	3	4	5
2	**Transparency & Explainability**		**This attribute measures the degree of transparency in the AI in terms of the algorithms, and models that comprise the AI & ML.**					
	Explanability Governance	1	To what degree are there formal policies or processes in place to govern explanability?	1	2	3	4	5
	Regulatory Sandbox (Visibility)	2	To what degree are there capabilities in place to provide visibility to regulators on explainability?	1	2	3	4	5
	Model Audit Controls	3	To what degree are there capabilities in place to audit models?	1	2	3	4	5
	Model Drift Prevention Monitoring	4	To what degree are there capabilities or processes in place to test & mitigate model drift?	1	2	3	4	5
	Use of Knowledge Graphs	5	To what degree are knowledge graphs leveraged to provide model explainability?	1	2	3	4	5
	Use of Model Card Reporting	6	To what degree are model cards leveraged to provide a decriptive model explainability?	1	2	3	4	5
	Advanced ML Explain (LIME, SHAP)	7	To what degree is there use of advanced black box technology such as SHAP or LIME?	1	2	3	4	5

Table 2: Survey Instrument, cont.

#	Factor Name	#	Key Evidence	Scoring				
3	**Fairness / Bias Mitigation**		**This attribute measures the ability to mitigate bias in lending and prevent discrimination harm.**					
	Policy for Fairness in Models	1	To what degree is there a policy in place to define the fairness rules in the models?	1	2	3	4	5
	Fairness in Training Data	2	To what degree are there fairness considerations in place in the training data?	1	2	3	4	5
	Human in the Loop	3	To what degree are there HITL (Human in the Loop) checkpoints in the ML workflow?	1	2	3	4	5
	Legal Implications	4	To what degree are you aware that there may be legal/compliance implications?	1	2	3	4	5
	Proxy Discrimination	5	To what degree does the capability to mitigate proxy discrimination exist?	1	2	3	4	5
	Model Reparation	6	To what extent are there processes to cure the bias if it is indeed found in the models?	1	2	3	4	5
4	**Data Management & Quality**		**This attribute measures the maturity of the data mgmt processes that feeds the ML models.**					
	Data Privacy	1	To what degree are there data privacy considerations to protect PII in compliance with GDPR?	1	2	3	4	5
	Differential Privacy Capability	2	To what degree are there differential privacy capability to protect PII in place?	1	2	3	4	5
	EDA for pre-modeling	3	To what degree do exploratory data analysis processes exist as part of pre-modelling?	1	2	3	4	5
	Use of Data Pipeline Tools	4	To what degree are there data pipeline tools in use for the ML?	1	2	3	4	5
	Use of Big Data Lake	5	To what degree is there a modernized big data lake environment in place?	1	2	3	4	5
	CDO involvement	6	To what degree is the CDO (Chief Data Officer) intimately involved with model risk management?	1	2	3	4	5
	Use of Synthetic Data	7	To what degree does the capability exist to supplement data with synthetic data?	1	2	3	4	5
	Use of DataOps	8	To what degree is there a DataOps process in place for collaborative data management?	1	2	3	4	5
	Right to be forgotten	9	To what degree is there a process in place to delete data for those who wish to be forgotten??	1	2	3	4	5
5	**Security**		**This attribute measures what specific data privacy and data security provisions are in place?**					
	Adversarial Cyber Attack Defense	1	To what degree are there ML adversarial attack defenses in place?	1	2	3	4	5
	Data Ecryption	2	To what degree are there data encryption provisions in place to secure the data?	1	2	3	4	5
	Special Security access for prod	3	To what extent is there a special level of security access to interact with the production models?	1	2	3	4	5
	Security Processes for Unintended	4	To what degree are there security processes in place to prevent unintended use of AI?	1	2	3	4	5
	Ability to disable Algos	5	To what degree are there controls in place to disable the algo if there is an issue with it?	1	2	3	4	5

Table 3: Pre-Validation Survey

Maturity of Responsible AI in Banking Instrument Validation Survey								
#	Factor Name	Key Evidence	Scoring				
1	**Categories**	To what extent do you agree that the survey instrument contains the relevant categories?	1	2	3	4	5
2	**Org Commitment attributes**	To what extent do you agree that the relevant attributes are included in the Org Commitment category?	1	2	3	4	5
3	**Transparency attributes**	To what extent do you agree that the relevant attributes are included in the Transparency category?	1	2	3	4	5
4	**Fairness attributes**	To what extent do you agree that the relevant attributes are included in the Fairness category?	1	2	3	4	5
5	**Data Management attributes**	To what extent do you agree that the relevant attributes are included in the Data Mgmt. category?	1	2	3	4	5
6	**Security attributes**	To what extent do you agree that the relevant attributes are included in the Security category?	1	2	3	4	5
7	**Category weightings**	To what extent to you agree that there should be even Weightings applied to the categories?	1	2	3	4	5
8	**Likert scale**	To what extent do you agree that the 5-point Likert scale is a good model for data collection?	1	2	3	4	5
9	**Survey Representation**	To what extent do you agree that the answers to these questions represent the elements of Mature RAI?	1	2	3	4	5
	Total score		**9**	**18**	**27**	**36**	**45**

Table 4: ESG Survey

Maturity of Responsible AI in Banking ESG Survey								
#	Factor Name	Key Evidence	Scoring				
1	**CSR/ESG - Executive Focus**	To what degree is there an executive strategic focus on CSR/ESG?	1	2	3	4	5
2	**CSR/ESG - Culture**	To what degree is there a culture of ESG present in the Bank?	1	2	3	4	5
3	**CSR/ESG - Training**	To what degree is there formal training required for ESG at the Bank?	1	2	3	4	5
4	**CSR/ESG - Environment**	To what degree is there an organizational focus on business impact to the environment?	1	2	3	4	5
5	**CSR/ESG - Social**	To what degree is there an organizational focus on the Bank's role in social issues?	1	2	3	4	5
6	**CSR/ESG - Governance**	To what degree is there an organizational focus on governance / compliance within the Bank?	1	2	3	4	5
	CSR/ESG Total Average		**6**	**12**	**18**	**24**	**30**

Table 5: Jobin Compendium

The table below is one of the most comprehensive early research efforts on the RAI principles and contains many of the sample principles that are incorporated into the survey instrument.

RAI Principle	# of documents present (84)	% of documents present?	Terms included in document search for coding.
Transparency	73	87%	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & Fairness	68	81%	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60	71%	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsiblility	60	71%	Responsibility, accountability, liability, acting with integrity
Privacy	47	56%	Privacy, personal or private information
Beneficience	41	49%	Benefits, beneficence, well-being, peace, social good, common good
Freedom & Autonomy	34	40%	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28	33%	Trust
Sustainability	14	17%	Sustainability, environment (nature), energy, resources (energy)
Dignity	13	15%	Dignity
Solidarity	6	7%	Solidarity, social security, cohesion

Source: Jobin – The Global Landscape of AI Ethics Guidelines (Jobin et al., 2019).

Table 6: Risk Assessment Frameworks & Fairness Toolkits

Below is a table of some of the popular fairness assessment and toolkits used in reviewing RAI.

#	Tool	Company	Reference	Type
1	Aequitas Bias and Fairness Audit Toolkit	Aequitas	http://aequitas.dssg.io/	Website Software Tool
2	AI Explainability 360 Open Source Toolkit	IBM	http://aix360.mybluemix.net	Website
3	AI Fairness 360 Open Source Toolkit by IBM	IBM	developer.ibm.com/articles/the-ai-360-toolkit-ai-models-explained/	GitHub Code
4	Algorithmic Accountability Policy Toolkit	AI Now Institute	https://ainowinstitute.org/aap-toolkit.pdf	White Paper
5	BlackBox Auditing Tool	Algofairness	https://github.com/algofairness/BlackBoxAuditing	GitHub Code
6	Deon	Deon	https://deon.drivendata.org/	Website
7	Fairness Flow	Facebook	http://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/	Website
8	Fairness Tool	Accenture	https://www.accenture.com/us-en/blogs/blogs-careers/were-harnessing-the-power-of-ai-to-benefit-all	Proprietary Code
9	audit-AI	Pymetriics	https://github.com/pymetrics/audit-ai	GitHub Code
10	Ethics & Algorithms Toolkit	Ethicstoolkit.ai	https://ethicstoolkit.ai/	Checklists
11	PWC Responsible AI Toolkit	PWC	https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html	Proprietary Code
12	InterpretML	Microsoft	https://github.com/interpretml/interpret	GitHub Code
13	FairLearn	Microsoft	https://github.com/fairlearn/fairlearn	GitHub Code
14	What-If Tool	Google (PAIR)	https://pair-code.github.io/what-if-tool/ai-fairness.html	GitHub Code

Table 7: Algorithms

Below is a table of some of the popular algorithms used in credit risk:

Algorithm Type	Algorithm Name	Description
Supervised	**Linear Regression**	Linear regression is a **supervised** learning algorithm and tries to model the relationship between a **continuous** target variable and one or more independent variables by fitting a linear equation to the data.
Supervised	**Support Vector Machine**	Support Vector Machine (SVM) is a **supervised** learning algorithm and mostly used for classification tasks, but it is also suitable for regression tasks.
Supervised	**Naïve Bayes**	Naive Bayes is a **supervised** learning algorithm used for classification tasks. Hence, it is also called Naive Bayes Classifier.
Supervised	**Logistic Regression**	Logistic regression is a **supervised** learning algorithm which is mostly used for **binary classification** problems.
Supervised	**Decision Tree**	A decision tree builds upon iteratively asking questions to partition data. It is easier to conceptualize the partitioning data with a visual representation of a decision tree
Supervised	**Random Forest**	Random forest is an ensemble of many decision trees. Random forests are built using a method called **bagging** in which decision trees are used as parallel estimators.
Supervised	**Gradient Boosted Decision Tree**	GBDT is an ensemble algorithm which uses **boosting** method to combine individual decision trees.
Supervised	**K-nearest neighbor (kNN)**	K-nearest neighbors (kNN) is a **supervised** learning algorithm that can be used to solve both classification and regression tasks. The main idea behind kNN is that the value or class of a data point is determined by the data points around it.
Unsupervised	**K-Means Clustering**	Clustering is a way to group a set of data points in a way that similar data points are grouped together.
Unsupervised	**Principal Component Analysis**	PCA is a dimensionality reduction algorithm which basically derives new features from the existing ones with keeping as much information as possible.

Source: https://towardsdatascience.com/

CHAPTER 3

RESEARCH METHODS

3.1 Description

This study is focused on assessing the maturity of the RAI program in Banks and measuring the related impact on the various stakeholder and shareholder attributes of the firm. In the research for this study, there was neither an accepted standard nor scoring index found for measuring the maturity of the RAI (MRAI) program, but rather several sets of principles (Benjamins et al., 2019; Boza & Evgeniou, 2021; Buhmann & Fieseler, 2021; Clarke, 2019; de Laat, 2021; Eitel-Porter, 2020; EUCommission, 2019; Fjeld et al., 2020; Floridi & Cowls, 2019; Floridi et al., 2018; Hagendorff, 2020; Institute, 2022; Jobin et al., 2019; Leslie, 2019; Morley et al., 2019; Myers & Nejkov, 2020; Rakova et al., 2020; Schiff et al., 2020; Siau & Wang, 2020; Zeng et al., 2019) as well as a few general self-assessment frameworks (Ayling & Chapman, 2021; Boza & Evgeniou, 2021; EUCommission, 2019; Mills & Duranton, 2021) that were published and available. There were also some maturity models (Vakkuri et al., 2020) and risk assessments (Fifth-Quadrant, 2021) in addition to initial pilot instruments (Coates & Martin, 2019) which this study can build upon. With an understanding of the current state of the academic body of knowledge regarding scoring standards for assessing MRAI, this study will create a novel instrument called "MRAI" and then use this tool for measuring MRAI in a 5-point Likert scale in 50 of the top US Banks within the study. In addition to the survey instrument development, there will be 2 methods of administration to assess the MRAI score, which will be comprised of a proxy MRAI score, and administering the MRAI survey instrument to the population of Banks as a guided self-assessment survey. As described in the instrument development section, this survey instrument will be comprised of 5 categories derived from an extensive analysis of common RAI principles. Each of the categories will have a set of sub-attributes, which will be assessed in a 5-point Likert scale, and cumulatively add up to an MRAI score for each Bank.

3.2 Data

3.2.1 Sample Description

The target sample population for the survey will be the largest banks in the United States. A large bank is defined as a bank that has over $25B in AUM (Assets under Management). The study aims to survey the population of 50 of the top US Banks and collect survey data for statistical analysis. The surveys will be conducted in a live interview over the phone or video-conference call. The rationale for this population is that these banks represent a significant portion of the credit lending for credit cards, mortgages, and auto loans in the United States. The Banks also have ample resources to employ mature RAI. These Banks are also public companies and highly regulated by the Fed and OCC, and as such, are more subject to social pressures for fairness and will provide a good status on the RAI capabilities built into their credit lending processes.

The study approach will have three components. First, there will be an interview to provide an extra validation step on the pre-validation survey instrument to ensure construct, face, and content validity of the instrument. If any of the assumed relevant categories and attributes of the survey instrument are deemed irrelevant or if there are missing elements, then the feedback may be incorporated into the final post-validation survey instrument. Once the instrument is validated, then the instrument will be leveraged to conduct the MRAI capability interview with each of the Banks. Since the survey instrument is focused on deriving an MRAI score, but likely has some related company attributes or traits that would be interesting to record as well, the third part of data collection of the study will further employ MTMM (Multi-Trait, Mono-Method) to ask some additional questions about implications of RAI on the ESG strategy. The three parts of the interview will be conducted at one setting lasting about 30 minutes. The instrument validation data, the MRAI capability data, and the ESG data will all be captured into their own survey constructs for specific analysis with Cronbach's alpha as well as CFA (Confirmatory Factor Analysis) to ensure internal consistency reliability.

3.2.1 Data Collection

The data will be collected in early 2022 for the MRAI score in two primary forms (MRAI Survey Instrument and MRAI Proxy Score). The study will also collect feedback on the actual instrument structure as well as data on the ESG strategy for the Banks.

3.2.1.1 MRAI Survey Instrument

First, is the survey instrument that was developed specifically for this study. As discussed above, the survey will be administered to 50 of the top US Banks with an interview of an MRM (Model Risk Management) executive that can provide a response to the Likert criteria for the questions in the survey. The attributes of the survey are listed in **Table 2**. The categories and the detailed attributes for the survey instrument will have performed the pre-validation step to provide face and content validity such that the post-validation instrument is robust, valid, and measuring the intended items to derive the study outcomes. The pre-validation survey interview is depicted in **Table 3**. As an additional component to the survey on MRAI, during the same interview session, the Banks will be asked about the ESG strategy (questions depicted in **Table 4**), for which the data to be incorporated into the MTMM statistical analysis.

3.2.1.2 MRAI Proxy Score

Second, is the MRAI proxy score data, which will be collected from a primary research effort scouring websites and the internet for information related to providing evidence of existence of the attributes of relevance for maturity of RAI capability.

These questions are detailed below and will be validated via the inter-rater reliability process that is described in the reliability section below.

1) Does an in-house research team exist for R(AI) and/or are there research publications available?
2) Are there R(AI) articles in the press published from the company?
3) Are there a set of published R(AI) principles available for the company?
4) Is R(AI) mentioned in the corporate documents (e.g., annual report)
5) Is there a direct link from the website to the R(AI) perspective or philosophy?

6) Are there published academic research partnerships for the company?

7) Does a R(AI) or related capability COE (center of excellence) exist for the company?

8) Are there published careers within the R(AI) field available for the company?

3.2.1.3 Confidentiality

This study has developed a coding method to anonymize the data from anyone except the key researcher. This confidentiality is consistent with guidance from the IRB (Institutional Review Board). This decoding will be performed by managing an offline spreadsheet that will correspond to a number (1-50) with a Bank name. In addition to prevent anyone from trying to guess the order of the Banks, this study has also randomized the numbers of the Banks. The Bank name will never become known outside the interviewer and the interviewee for the primary survey instrument and kept confidential in the secondary proxy MRAI sheet.

3.3 Validity and Reliability

3.3.1 Validity

With the goal of development of a robust MRAI survey instrument and employing MTMM (Multi-Trait Multi-Method) (Campbell & Fiske, 1959), this study will capture the survey data for the Banks for the RAI principles categories as well as record Bank information related to the ESG-CSR (Environmental, Social, Governance - Corporate Social Responsibility), which may provide some additional insights about the linkage between RAI and ESG-CSR.

3.3.1.1 Face Validity

In terms of face validity, the survey attributes will be validated by the pre-validation instrument interviews, which will review if the questions being asked and attributes deriving the survey, appear to be relevant to the field of study. Another test of face validity is the extensive compendium shown in **Table 1**, which provides credibility to the components of the instrument.

3.3.1.2 Content Validity

Following a similar process during the pre-validation instrument interviews as the test for face validity, this study will also be concerned with whether the content validity pertains to the specific scientific context of the AI/ML models that are used in credit lending and measures the maturity of the RAI capability, which will be validated by the initial set of interviews. As previously indicated, if there are questions or attributes in the survey that are irrelevant or it is deemed that there are critical components missing, they will be identified in the pre-validation instrument interviews and incorporated into the post-validation survey instrument.

3.3.1.3 Construct Validity

The study leveraged the MTMM framework (Campbell & Fiske, 1959) to test for the instrument's construct validity. This study will test the MRAI measure for both convergent and discriminant validity. To test for convergence and discriminant validity, this study will correlate the MRAI score with two ESG scores (Instrument ESG and Sustainalytics ESG score). To satisfy this requirement of the study for convergent validity, the instrument must be significantly correlated with a conceptually similar construct (such as the Proxy MRAI score). In the case of discriminant validity, the variables must *not* be significantly correlated with a seemingly related but conceptually different construct (e.g., Sustainalytics ESG score). In this case, the test is for a similar method, but different traits, which is interviewing the Bank executives for both MRAI (using the newly developed survey) as well as the ESG-CSR scoring (using the ESG instrument) in the same interview method. Conversely, if the same trait is evaluated, but different methods involving the MRAI instrument are used, but then compared with an MRAI proxy, which is collected not by interview, but instead by primary research on publicly available information and proxy interpreted scoring will be employed. In the case described above, in fact, if MRAI scores correlate higher with the Instrument ESG score than with Sustainalytics ESG, the study will have some evidence of both convergent and discriminant validity. In terms of relating an ESG-CSR metric with a reputation index and testing with the MTMM method, Rahman, et al. (Rahman & Blake, 2021) set a precedent in testing for convergent and discriminant validity.

3.3.2 Reliability

3.3.2.1 Internal Consistency Reliability

Since each of the RAI principles categories has a different number of sub-elements, this study will utilize a couple of techniques to ensure that the most important elements are included. First, to ensure the internal consistency reliability, this study will leverage Cronbach's alpha which is used in instruments where Likert questions are present and correlates the relatedness between the survey questions. This study will aim for each of the elements having a Cronbach's alpha of ($\alpha > 0.7$), or the question can be considered for removal to ensure maximum reliability. In addition, this study will employ CFA (confirmatory factor analysis) to test for internal consistency reliability and ensure that the factors of the survey instrument are statistically meaningful and without outliers or other issues, such as collinearity. In the event of finding one or more of the elements of the survey instrument, this study may consider removing the factor from the instrument. The CFA scores will aim to be ($c > .5$).

3.3.2.2 Inter-rater Reliability

As described in the above section, this study will employ a Proxy MRAI score in addition to the MRAI instrument survey interview self-assessment, such that this deployment yields data for MTMM (Mono-Trait, Multi-Method). This study will leverage two trained coders to interpret the data for evidence of the MRAI proxy questions and record 1, .5, or 0 if evidence is found for the attribute. As a reliability test, this study will employ an inter-rater reliability test for the proxy MRAI score, which will be comprised of 400 items (8 items x 50 Banks) as there is a review and judgement to be conducted for each attribute on whether the evidence is true or false in meeting the criteria to record a 1, .5, or 0 accordingly. The specific references that derived this researcher's score will be documented so that the coding can be reviewed or reperformed by another rater to provide the inter-rater reliability attestation. This study will leverage the weighted Cohen's Kappa coefficient (measures the degree of agreement between the 2 raters) (Cohen, 1960) to ensure that the reliability is above 0.6 and within ($p<.01$) to account for any impact due to the chance correlation effect (Landis & Koch, 1977).

3.4 Summary

The measurement of RAI is immature and contains a gap that is being addressed with various inconsistent and uncoordinated methods. There are many RAI principles available from different companies as well as consultancies and cooperative groups that seem to be converging on a standard. There are some preliminary instruments (Coates & Martin, 2019), assessment frameworks, toolkits (AIEthicist, 2021; Benjamins, 2020; Boza & Evgeniou, 2021), certificate programs (Institute, 2022) and maturity models (Sadiq et al., 2021; Vakkuri et al., 2021; Vakkuri et al., 2020) that are progressively being adopted in the industry.

This study has conducted a comprehensive review of the current state of the RAI principles and supporting assessment capabilities that incorporate the most common elements into the proposed MRAI survey instrument for credit lending. Through the robust diligence in developing the instrument as well as demonstrating the validity and reliability, this tool may be an advancement for the industry to leverage in governing the responsibility of AI. With growing momentum around pending regulation (Almeida et al., 2021; Burt, 2021; Candelon et al., 2021; Crosman, 2022; Wyden et al., 2022) to require Banks more formally by law to comply with this new governance, a standard tool will be a significant contribution to the field.

Through the introduction of a new standard MRAI survey instrument, this study will conduct data collection through two different methods, first with the novel MRAI survey instrument, then as a supplement with a MRAI proxy analysis. This study will then run extensive statistical analysis on each set of data and attempt to illuminate the current state of maturity of RAI for credit lending within Banks.

CHAPTER 4

RESULTS

4.1 Overview

This study employed multiple reliability and validity methods to provide support for the robustness and integrity of the research on the data of the 48 Banks interviewed. The statistical methods of Cronbach's alpha, CFA (Confirmatory Factor Analysis), and Inter-Rater reliability using Cohen's Kappa all provide support for internal consistency reliability. MTMM (Multi-Trait-Multi-Method) (Campbell and Fiske, 1959) provides support for construct validity in measuring both convergent and discriminant validity for the MRAI trait. The various data sets that were collected as part of the study had the specific statistical methods applied, and the explanation of the results is below.

There were multiple aspects involved with this study in developing, validating, and administering the novel MRAI instrument. First, this study performed comprehensive research on the common RAI principles (**Table 1**) and selected the relevant categories (Organizational Commitment, Explainability, Fairness, Data Mgmt., Security) for the pre-validation instrument based on the most cited RAI principles in the related literature. In addition to the high-level instrument 'categories' employed, which were supported by the aforementioned research, this study also leveraged additional extant literature to support each 'attribute' to provide a more granular level of instrumentation resulting in the detailed attributes within each category. The attributes ranged from five to nine elements per instrument category. This aspect of the instrument (incorporating both categories and detailed attributes) is important as the study demonstrates the internal consistency reliability robustness of the instrument through statistical evidence (using Cronbach's alpha and CFA) of these attributes in relation to the particular category.

Once the draft instrument (this study named it the pre-validation instrument) was developed, the study then conducted interviews to perform face and content validity reviews for each of the categories and attributes. The study captured this feedback data in a separate interview survey and conducted a

Cronbach's alpha as well as CFA statistical analysis on this data. Analyzing both the pre-validation instrument, as well collecting the Bank data with the instrument then led to two sets of Cronbach's alpha and CFA internal consistency reliability statistical analyses. The first set of statistics measured the reliability of the pre-validated instrument design categories. The second set of statistics measured the consistency of the attribute elements of the instrument by measuring the data collected from the Bank's with the post-validation instrument. In addition to these two statistics tests, this study ran another additional test (which was conducted at the same time as the MRAI capability interview) surveying questions related to the Bank's ESG strategy. Cronbach's alpha as well as CFA analysis were performed on the ESG data collection and the general correlations to MRAI were used for the MTMM analysis.

4.2 Instrument Reliability – Pre-Validation analysis

As a first validation step in the instrument development process, this study interviewed 48 bank executives in the MRM area to collect feedback on the relevance of the categories and attributes that comprised the elements of the instrument. The study computed Cronbach's alpha for the overall MRAI instrument and was interested in the categories, relevancy of the attributes, the potential weightings, the Likert scale, and the overall relevance factor. In addition to Cronbach's alpha and CFA quantitative analysis, the interview feedback highlighted some qualitative items as well. First was a focus on the Organizational Commitment category with a comment that the "Investment" and "ROI elements" were similar. Second was a comment on the "Pareto element" in that it was not being utilized by Banks pervasively. Thirdly, there was a point about the "Data quality CDO" involvement question in terms of highlighting that could have been in the Organizational Commitment category. Lastly there was a comment about the Likert scale and a question on if the study considered using a binary scale of 0 or 1 for measuring existence of the capability for each of the elements. With this feedback from a few interviewees taken into consideration, this study decided to continue with the original instrument as the data did not overwhelmingly support adding or removing particular elements; and there was extant literature supporting these important elements in the specific categories.

The study performed Cronbach's alpha on the pre-validation tool elements to ensure internal consistency reliability of the instrument categories relevance and the score was 0.889.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.889	.900	9

Item Statistics

	Mean	Std. Deviation	N
Cat	4.50	.652	48
OC	4.58	.577	48
Explain	4.56	.580	48
Fair	4.56	.580	48
Data	4.65	.526	48
Security	4.65	.565	48
Weight	4.63	.606	48
Likert	4.65	.565	48
Relevance	4.38	.703	48

Inter-Item Correlation Matrix

	Cat	OC	Explain	Fair	Data	Security	Weight	Likert	Relevance
Cat	1.000	.226	.422	.366	.155	.202	.108	.087	.232
OC	.226	1.000	.588	.778	.766	.843	.639	.713	.341
Explain	.422	.588	1.000	.747	.807	.621	.613	.491	.098
Fair	.366	.778	.747	1.000	.737	.816	.734	.686	.150
Data	.155	.766	.807	.737	1.000	.787	.777	.644	.194
Security	.202	.843	.621	.816	.787	1.000	.661	.733	.234
Weight	.108	.639	.613	.734	.777	.661	1.000	.599	.187
Likert	.087	.713	.491	.686	.644	.733	.599	1.000	.181
Relevance	.232	.341	.098	.150	.194	.234	.187	.181	1.000

Cronbach's alpha statistics provide significant support for the internal consistency reliability for the element of the instrument, the scale being used, as well as the weighting and overall relevance of the instrument.

4.3 Instrument Reliability - CFA Analysis on Pre-Validation elements

This study also performed a CFA internal consistency test in addition to the Cronbach's alpha with the results below. The CFA absolute value suppression threshold was set to .5, and as is depicted in the first statistics run on the left side below, the 'Weight' attribute was .444, and below the .5 threshold. The 'Weight' element was removed for the subsequent statistics run which is depicted on the right below, with some minor changes to the remaining elements.

Statistics with the "Weight" element included.

Communalities

	Initial	Extraction
Cat	1.000	.714
OC	1.000	.829
Explain	1.000	.737
Fair	1.000	.808
Data	1.000	.666
Security	1.000	.665
Weight	1.000	.444
Likert	1.000	.616
Relevance	1.000	.613

Extraction Method: Principal Component Analysis.

Statistics with the "Weight" element removed

Communalities

	Initial	Extraction
Cat	1.000	.775
OC	1.000	.842
Explain	1.000	.748
Fair	1.000	.794
Data	1.000	.702
Security	1.000	.685
Likert	1.000	.607
Relevance	1.000	.640

Extraction Method: Principal Component Analysis.

One key finding to highlight was that as a result of the "Weight" element being removed from the CFA, the study maintained the same weighting for all of the categories and attribute elements in the post-validation survey instrument. In addition to providing qualitative support for the face and content validity from the model risk management executives, the study's statistics illustrate quantitatively that there is internal consistency reliability for the instrument structure, categories, as well as detailed attributes.

4.4 Post-Validation Instrument Reliability

With the instrument validation confirmed regarding face and content as well as internal consistency reliability confirmation from Cronbach's alpha test as well as the CFA statistics, this study leveraged the instrument to interview 48 bank executives in the MRM (model risk management) area to collect data regarding capabilities for the Banks. This study then used these scores to perform an additional set of internal consistency reliability validations with Cronbach's alpha and CFA. The study computed Cronbach's alpha for the attribute items within each of the five categories (Organizational Commitment, Explainability, Fairness, Data Quality, and Security) of the MRAI instrument.

4.4.1 Cronbach's alpha – Organizational Commitment

Cronbach's alpha for the seven Organizational Commitment items was 0.898 providing significant statistical support for the internal consistency reliability.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.898	.899	7

Item Statistics

	Mean	Std. Deviation	N
OC - Org	3.00	1.052	48
OC - Invest	3.04	.944	48
OC - ROI	2.65	1.101	48
OC - Train	2.92	.895	48
OC - Culture	3.06	1.040	48
OC - Pareto	1.75	1.000	48
OC - C-Suite	2.90	1.016	48

Inter-Item Correlation Matrix

	OC - Org	OC - Invest	OC - ROI	OC - Train	OC - Culture	OC - Pareto	OC - C-Suite
OC - Org	1.000	.707	.680	.587	.623	.182	.637
OC - Invest	.707	1.000	.649	.734	.691	.349	.692
OC - ROI	.680	.649	1.000	.509	.726	.285	.670
OC - Train	.587	.734	.509	1.000	.554	.190	.669
OC - Culture	.623	.691	.726	.554	1.000	.404	.792
OC - Pareto	.182	.349	.285	.190	.404	1.000	.414
OC - C-Suite	.637	.692	.670	.669	.792	.414	1.000

4.4.2 Cronbach's alpha – Explainability

Cronbach's alpha for the seven Explainability items was 0.951, providing significant statistical support for the internal consistency reliability.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.951	.951	7

Item Statistics

	Mean	Std. Deviation	N
E - Gov	3.42	1.235	48
E - Reg	3.21	1.148	48
E - Audit	3.19	1.197	48
E - Drift	3.06	1.295	48
E - Graph	2.79	1.166	48
E - Card	2.83	1.173	48
E - LIME	2.02	1.211	48

Inter-Item Correlation Matrix

	E - Gov	E - Reg	E - Audit	E - Drift	E - Graph	E - Card	E - LIME
E - Gov	1.000	.823	.882	.808	.741	.754	.492
E - Reg	.823	1.000	.792	.821	.717	.753	.578
E - Audit	.882	.792	1.000	.871	.806	.735	.599
E - Drift	.808	.821	.871	1.000	.812	.805	.637
E - Graph	.741	.717	.806	.812	1.000	.830	.606
E - Card	.754	.753	.735	.805	.830	1.000	.586
E - LIME	.492	.578	.599	.637	.606	.586	1.000

4.4.3 Cronbach's alpha – Fairness

Cronbach alpha for the six Fairness items was 0.931, providing significant statistical support for the internal consistency reliability.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.931	.932	6

Item Statistics

	Mean	Std. Deviation	N
F - Policy	3.50	1.220	48
F - Data	3.25	1.296	48
F - Human	3.17	1.226	48
F - Legal	3.02	1.082	48
F - Proxy	2.90	1.016	48
F - Repair	2.42	1.182	48

Inter-Item Correlation Matrix

	F - Policy	F - Data	F - Human	F - Legal	F - Proxy	F - Repair
F - Policy	1.000	.888	.796	.637	.678	.501
F - Data	.888	1.000	.830	.740	.715	.555
F - Human	.796	.830	1.000	.687	.817	.582
F - Legal	.637	.740	.687	1.000	.680	.675
F - Proxy	.678	.715	.817	.680	1.000	.657
F - Repair	.501	.555	.582	.675	.657	1.000

4.4.4 Cronbach's alpha – Data Quality

Cronbach's alpha for the nine Data Quality items was 0.947, providing significant statistical support for the internal consistency reliability.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.947	.948	9

Item Statistics

	Mean	Std. Deviation	N
D - Privacy	3.96	1.237	48
D - Differ	3.27	1.162	48
D - EDA	3.04	1.237	48
D - Pipe	3.17	1.038	48
D - Big	3.21	1.129	48
D - CDO	3.06	1.040	48
D - Synthetic	2.10	1.134	48
D - Ops	2.60	1.106	48
D - Forget	2.63	1.160	48

Inter-Item Correlation Matrix

	D - Privacy	D - Differ	D - EDA	D - Pipe	D - Big	D - CDO	D - Synthetic	D - Ops	D - Forget
D - Privacy	1.000	.748	.752	.685	.661	.581	.428	.532	.567
D - Differ	.748	1.000	.717	.755	.686	.584	.624	.698	.661
D - EDA	.752	.717	1.000	.773	.816	.627	.543	.681	.678
D - Pipe	.685	.755	.773	1.000	.805	.680	.599	.818	.689
D - Big	.661	.686	.816	.805	1.000	.714	.581	.766	.776
D - CDO	.581	.584	.627	.680	.714	1.000	.536	.688	.567
D - Synthetic	.428	.624	.543	.599	.581	.536	1.000	.644	.677
D - Ops	.532	.698	.681	.818	.766	.688	.644	1.000	.728
D - Forget	.567	.661	.678	.689	.776	.567	.677	.728	1.000

4.4.5 Cronbach's alpha – Security

Cronbach's alpha for the five Security items was 0.892, providing significant statistical support for the internal consistency reliability.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.892	.891	5

Item Statistics

	Mean	Std. Deviation	N
S - Cyber	3.75	1.212	48
S - Encrypt	3.85	1.220	48
S - Access	3.19	1.232	48
S - Special	2.83	1.155	48
S - Disable	2.27	1.144	48

Inter-Item Correlation Matrix

	S - Cyber	S - Encrypt	S - Access	S - Special	S - Disable
S - Cyber	1.000	.709	.645	.487	.188
S - Encrypt	.709	1.000	.797	.707	.517
S - Access	.645	.797	1.000	.770	.613
S - Special	.487	.707	.770	1.000	.776
S - Disable	.188	.517	.613	.776	1.000

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
S - Cyber	12.15	17.446	.583	.608	.902
S - Encrypt	12.04	15.402	.828	.725	.846
S - Access	12.71	15.062	.862	.746	.838
S - Special	13.06	15.890	.825	.764	.848
S - Disable	13.63	17.771	.595	.673	.898

4.4.6 CFA Analysis – MRAI Instrument

This study computed the CFA for the entire MRAI instrument as well. The CFA for each of the instrument items was above .6, with the lowest value (Data - CDO) being .638 and the highest value (Fairness – Data Training) being .873. The results below provide statistical support for the internal consistency reliability.

OC - Org	1.000	.735
OC - Invest	1.000	.793
OC - ROI	1.000	.699
OC - Train	1.000	.696
OC - Culture	1.000	.800
OC - Pareto	1.000	.701
OC - C-Suite	1.000	.837
E - Gov	1.000	.836
E - Reg	1.000	.854
E - Audit	1.000	.862
E - Drift	1.000	.867
E - Graph	1.000	.802
E - Card	1.000	.834
E - LIME	1.000	.827
F - Policy	1.000	.854
F - Data	1.000	.873
F - Human	1.000	.811
F - Legal	1.000	.713
F - Proxy	1.000	.796
F - Repair	1.000	.738
D - Privacy	1.000	.828
D - Differ	1.000	.728
D - EDA	1.000	.753
D - Pipe	1.000	.796
D - Big	1.000	.854
D - CDO	1.000	.638
D - Synthetic	1.000	.799
D - Ops	1.000	.790
D - Forget	1.000	.741
S - Cyber	1.000	.836
S - Encrypt	1.000	.801
S - Access	1.000	.835
S - Special	1.000	.746
S - Disable	1.000	.841

These scores from both Cronbach's alpha as well as CFA statistics offer convincing evidence of internal consistency reliability of our MRAI measurement instrument.

4.5 Proxy MRAI & Inter-Rater Reliability

For the Proxy MRAI method, the research practitioner reviewed 48 Bank website sources, and scored eight elements for relationship with RAI capabilities (RAI research department, RAI articles, published RAI principles, RAI mention in the 10k, RAI link on website, university innovation RAI partnerships, RAI COE, careers in RAI). The scoring was coded as a 0 for no AI or RAI evidence present, a .5 for AI presence, but not RAI, and 1 for RAI presence. Since the instrument requires raters to evaluate various MRAI attributes of the Banks, examining inter-rater reliability is recommended. The initial inter-rater agreements resulted in 97.7% for the Banks, and even after accounting for chance correlation the kappa coefficients for our two different raters (Cohen, 1960), the kappa coefficient is 0.965 ($p < .001$) for the two raters who coded the 48 Banks. The Cohen kappa is high because the references for the coding were specifically recorded, aligning the agreement. These kappa coefficients fall well within the range of substantial agreement even considering change correlation (Landis and Koch, 1977). These results provide extremely strong evidence for inter-rater internal consistency reliability of the MRAI instrument.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Researcher Score * Assistant Score	384	100.0%	0	0.0%	384	100.0%

Researcher Score * Assistant Score Crosstabulation

		Assistant Score							
		.0		.5		1.0		Total	
		N	%	N	%	N	%	N	%
Researcher Score	.0	132	93.6%	0	0.0%	0	0.0%	132	34.4%
	.5	4	2.8%	131	100.0%	0	0.0%	135	35.2%
	1.0	5	3.5%	0	0.0%	112	100.0%	117	30.5%
Total		141	100.0%	131	100.0%	112	100.0%	384	100.0%

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.965	.012	26.715	<.001
N of Valid Cases		384			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

4.6 Instrument ESG

The third component of the Bank interviews was to conduct survey questions about a related but different item named ESG (Environment, Social, Governance). This additional Bank feedback provides the study with another ESG variable from a method that will enable the comparison of multi-trait mono-method, since these ESG questions were asked in the same interview session (again using a 5-point Likert scale) as the core instrument feedback, thus the same method, however, related, but different traits. Below are the attributes of the ESG survey and they are also depicted in **Table 4**.

As with the other instrument data for the Bank's, this study ran Cronbach's alpha (.877) and CFA (all factors > .5) to be consistent with our prior statistics regarding elements of the instrument.

Cronbach's alpha

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.877	.881	6

Item Statistics

	Mean	Std. Deviation	N
Exec	3.10	.951	48
Culture	3.29	.922	48
Training	3.42	.794	48
Env	3.31	1.223	48
Social	3.02	.978	48
Gov	3.10	.751	48

Inter-Item Correlation Matrix

	Exec	Culture	Training	Env	Social	Gov
Exec	1.000	.596	.448	.667	.478	.462
Culture	.596	1.000	.528	.597	.630	.509
Training	.448	.528	1.000	.586	.454	.461
Env	.667	.597	.586	1.000	.635	.497
Social	.478	.630	.454	.635	1.000	.721
Gov	.462	.509	.461	.497	.721	1.000

CFA (Confirmatory Factor Analysis)

Communalities

	Initial	Extraction
Exec	1.000	.589
Culture	1.000	.665
Training	1.000	.522
Env	1.000	.710
Social	1.000	.690
Gov	1.000	.588

Extraction Method: Principal Component Analysis.

4.7 Validity

4.7.1 Multi-Trait Multi-Method (MTMM)

This study leveraged MTMM (Campbell and Fiske, 1959) for construct validity testing and collected two different MRAI trait scores with differing methods (proxy method, and instrument method). This established the same trait in MRAI, but different methods to be calculated with regression for the MTMM analysis to test for convergent validity. The first method for collecting the MRAI was the proxy MRAI score, which was derived by primary research investigating public attributes on the Bank's websites, publications, and internet. The second method for collecting the MRAI was administering the actual MRAI instrument that was described above, with a method of survey interview with the relevant MRM Bank executives. This establishes the same MRAI trait, and a different method of data collection. The study tested for convergent validity by regressing the two MRAI traits, with the results of an ($r = .882$) with ($p < .001$) providing significant support for the convergent validity.

4.7.2 Convergent validity – Instrument MRAI with Proxy MRAI Correlation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.882[a]	.778	.773	9.99492%

a. Predictors: (Constant), V4

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16120.006	1	16120.006	161.364	<.001[b]
	Residual	4595.326	46	99.898		
	Total	20715.332	47			

a. Dependent Variable: V3

b. Predictors: (Constant), V4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-11.640	4.915		-2.368	.022	-21.534	-1.747
	V4	1.004	.079	.882	12.703	<.001	.845	1.164

4.7.3 MTMM Matrix:

In the MTMM matrix below, we display various correlations: MRAI scores (Instrument MRAI and Proxy MRAI) obtained from different methods (mono-trait multi-method), MRAI scores and ESG scores obtained from different methods (multi-trait multi-method), MRAI scores and ESG scores from the same method (multi-train mono-method) and ESG scores from different methods (mono-trait multi-method). With the goal being to test for construct validity of the proposed MRAI instrument, the study collected data from MRAI (Instrument and Proxy) measuring instruments and ESG measuring instruments (Instrument ESG and Sustainalytics). This allowed the study to compute multi-trait multi-method correlations. To contrast MRAI against a different construct, we used both an instrument scoring method and a proxy score method. This allowed us to compute mono-trait multi-method correlations. This study leverages Campbell and Fiske (Campbell & Fiske, 1959), which explains that the presence of construct validity (i.e., both convergent and discriminant validity) is observed if the following comparisons show success:

(i) The correlation derived for a given construct (i.e., mono-trait) but scored through two different instruments (i.e., multi-method) exceeds both (a) the correlation comparing varied constructs (i.e., multi-trait) assessed through the same instrument (i.e., mono-method) and (b) the correlation comparing different constructs (i.e., multi-trait) calculated through alternative instruments (i.e., multi-method).

(ii) The correlation derived for independent constructs (i.e., multi-trait) computed through the same instrument (i.e., mono-method) exceeds the correlation between varied constructs (i.e., multi-trait) scored through alternative instruments (i.e., multi-method).

In the matrix below, the study finds that the mono-trait multi-method correlation of MRAI is significantly higher. The correlation is (r=.882) between the proposed MRAI instrument and the Proxy MRAI. The study also computed both Instrument MRAI and Proxy MRAI, regressing these traits against a related, but different trait in two ESG index scores (Instrument ESG and Sustainalytics) (mono-trait multi-method). The correlations between both Instrument MRAI and Instrument ESG (r=.553) and Proxy MRAI and Instrument ESG (r=.398) were both significantly lower than the mono-trait correlation

between Instrument MRAI and proxy MRAI (r=.882) demonstrating discriminant validity. The correlations between both Instrument MRAI and Sustainalytics (r=.135) and Proxy MRAI and Sustainalytics (r=.109) were both significantly lower than both mono-trait correlations (Instrument MRAI and Proxy MRAI) comparing (r=.882) and the Instrument ESG correlations (r=.553) for Instrument MRAI, and (r=.398) for Proxy MRAI. Lastly the study was interested in a correlation between the two ESG methods (Instrument ESG and Sustainalytics), which resulted in a correlation of (r=.532), which also is significantly lower than the convergent validity with the MRAI regression (r=.882), again demonstrating an example of discriminant validity. Although the multi-trait, mono-method score for Instrument MRAI and Instrument ESG was lower than (r=.882) with (r=.532) it should be noted that it is higher than Proxy MRAI and Instrument ESG (r=.398) and Instrument MRAI and Sustainalytics (r=.135). Overall, the MTMM analysis displayed below demonstrates clear evidence of construct validity, highlighting both convergent as well as discriminant validity for the proposed MRAI survey instrument.

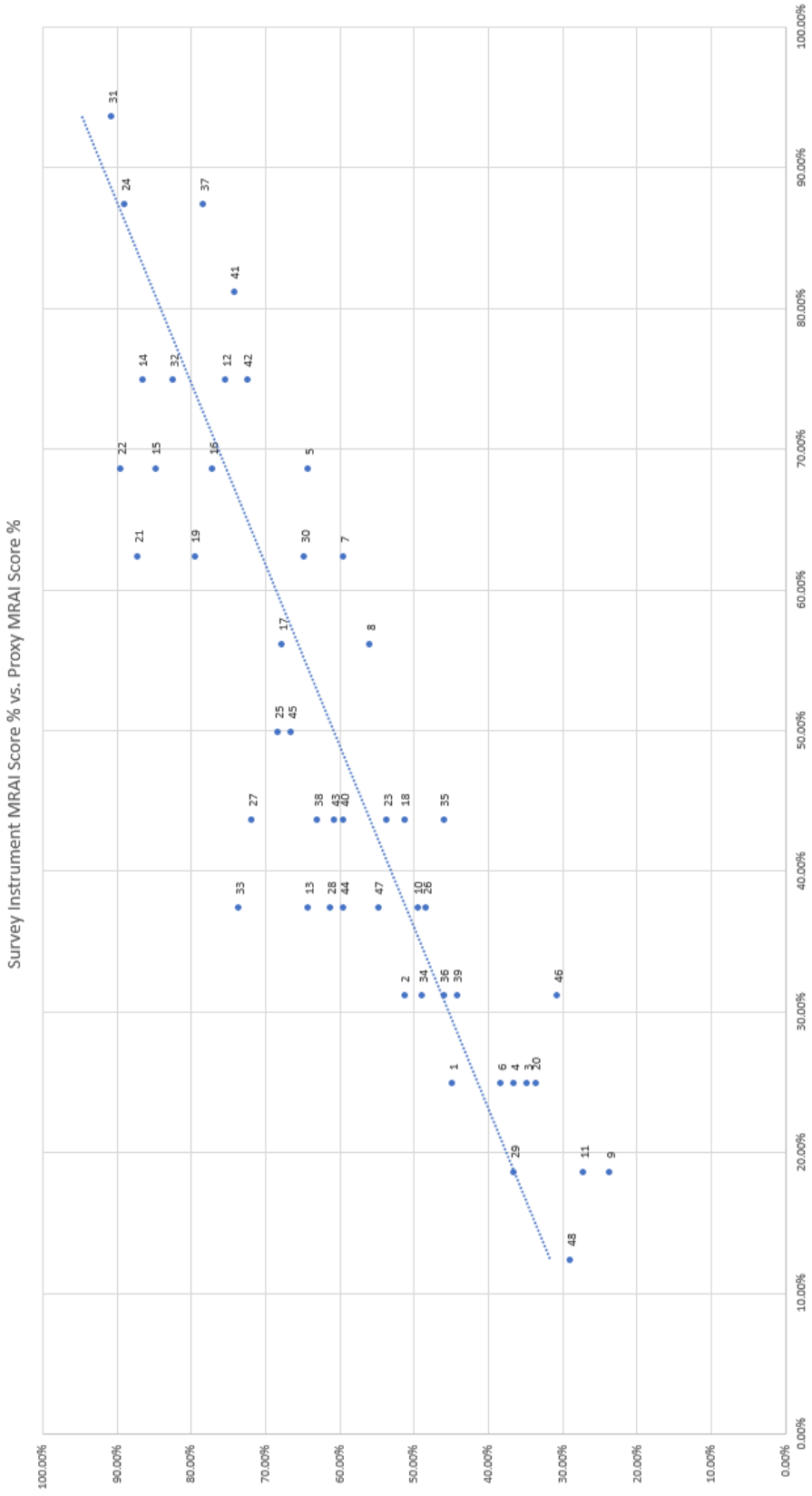
4.7.3.1 Table 8 - MTMM Matrix:

Instrument	Instrument MRAI	MTMM Type	Proxy MRAI	MTMM Type	Instrument ESG	MTMM Type
Proxy MRAI	0.882	mono-trait - multi-method				
Instrument ESG	0.553	multi-trait - mono-method	0.398	multi-trait - multi-method		
Sustainalytics ESG	0.135	multi-trait - multi-method	0.109	multi-trait - multi-method	0.532	mono-trait - multi-method

4.7.4 Instrument MRAI vs. Proxy MRAI correlation graph:

The graph in **Figure 7** depicts the data with the Instrument MRAI on the Y-axis and the Proxy MRAI on the X-axis. The data illustrates is what is expected in that a Bank's low Proxy MRAI score is typically highly correlated with a low Instrument MRAI score, and conversely that a Bank's high Proxy MRAI score is typically highly correlated with a high Instrument MRAI score. The Bank's are mapped to an anonymous number (1 - 48) as well as randomized guaranteeing robust confidentiality.

Figure 7: MRAI Instrument vs. MRAI Proxy correlation



4.7.5 Discriminant Validity

In order to analyze for discriminant validity, the study introduced a separate variable, creating a scenario where the study had the same method, but now a different trait. In order to address this requirement, the instrument must not be significantly correlated with a related but conceptually different construct. This new trait was the Bank's ESG score, of which the study collected two different ESG index scores (Instrument ESG and, Sustainalytics).

The (r=.553) for the correlation between Instrument MRAI and Instrument ESG score, which is significantly less than the MRAI (r = .882), provides support for discriminant validity.

Instrument MRAI – Instrument ESG Correlation:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.553[a]	.306	.291	13.18755%

a. Predictors: (Constant), Instrument_MRAI

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3530.739	1	3530.739	20.302	<.001[b]
	Residual	7999.928	46	173.911		
	Total	11530.667	47			

a. Dependent Variable: Instrument_ESG

b. Predictors: (Constant), Instrument_MRAI

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	36.399	6.485		5.613	<.001
	Instrument_MRAI	.470	.104	.553	4.506	<.001

a. Dependent Variable: Instrument_ESG

The (r=.398) for the correlation between Proxy MRAI and Instrument ESG score, which is significantly less than the MRAI (r = .882), provides support for discriminant validity.

4.7.5.1 Proxy MRAI – Instrument ESG Correlation:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.398[a]	.158	.140	19.468573%

a. Predictors: (Constant), Instrument_ESG

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3280.166	1	3280.166	8.654	.005[b]
	Residual	17435.166	46	379.025		
	Total	20715.332	47			

a. Dependent Variable: Proxy_MRAI

b. Predictors: (Constant), Instrument_ESG

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13.734	11.998		1.145	.258
	Instrument_ESG	.533	.181	.398	2.942	.005

a. Dependent Variable: Proxy_MRAI

In addition, to provide another example of discriminant validity, the study also tested an ESG score from the Sustainalytics source.

The (r=.135) for the correlation between Instrument MRAI and the Sustainalytics ESG score, which is significantly less than the MRAI (r = .882), provides support for discriminant validity.

4.7.5.2 Instrument MRAI - Sustainalytics Correlation:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.135[a]	.018	-.003	16.07997%

a. Predictors: (Constant), Survey %

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	222.101	1	222.101	.859	.359[b]
	Residual	11894.006	46	258.565		
	Total	12116.107	47			

a. Dependent Variable: Sustain

b. Predictors: (Constant), Survey %

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	62.171	7.908		7.862	<.001
	Survey %	.118	.127	.135	.927	.359

a. Dependent Variable: Sustain

The (r = .109) for the correlation between Proxy MRAI and the Sustainalytics ESG score, which is significantly less than the MRAI (r = .882), provides support for discriminant validity.

4.7.5.3 Proxy MRAI - Sustainalytics Correlation:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.109[a]	.012	-.010	16.13248%

a. Predictors: (Constant), Proxy %

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	144.285	1	144.285	.554	.460[b]
	Residual	11971.822	46	260.257		
	Total	12116.107	47			

a. Dependent Variable: Sustain

b. Predictors: (Constant), Proxy %

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	65.167	5.867		11.107	<.001
	Proxy %	.083	.112	.109	.745	.460

a. Dependent Variable: Sustain

The graph in **Figure 8** depicts the data with the Instrument MRAI on the Y-axis and the Sustainalytics data on the X-axis. The data illustrates the low correlation between the MRAI Instrument and the Sustainalytics data, which demonstrates the discriminant validity.

In addition to the MRAI regression analyses, in the context of MTMM, the study also performed a correlation between the ESG sources to determine the relationship. As expected, the (r = .532) for the correlation between Instrument ESG and the Sustainalytics ESG score is significantly less than the MRAI (r = .882) providing additional support for discriminant validity.

4.7.5.4 Instrument ESG - Sustainalytics Correlation:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.532[a]	.283	.267	13.40739%

a. Predictors: (Constant), Sustainalytics

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3261.795	1	3261.795	18.145	<.001[b]
	Residual	8268.871	46	179.758		
	Total	11530.667	47			

a. Dependent Variable: Instrument_ESG

b. Predictors: (Constant), Sustainalytics

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	28.441	8.645		3.290	.002
	Sustainalytics	.519	.122	.532	4.260	<.001

a. Dependent Variable: Instrument_ESG

The graph in **Figure 9** depicts the data with the Instrument MRAI on the Y-axis and the Instrument ESG data on the X-axis. The data illustrates the medium correlation between the MRAI Instrument and the Instrument ESG data, which demonstrates the discriminant validity with the MRAI correlation, but Instrument ESG is much higher than the Sustainalytics correlation.

Figure 8: MRAI Instrument vs. Sustainalytics correlation

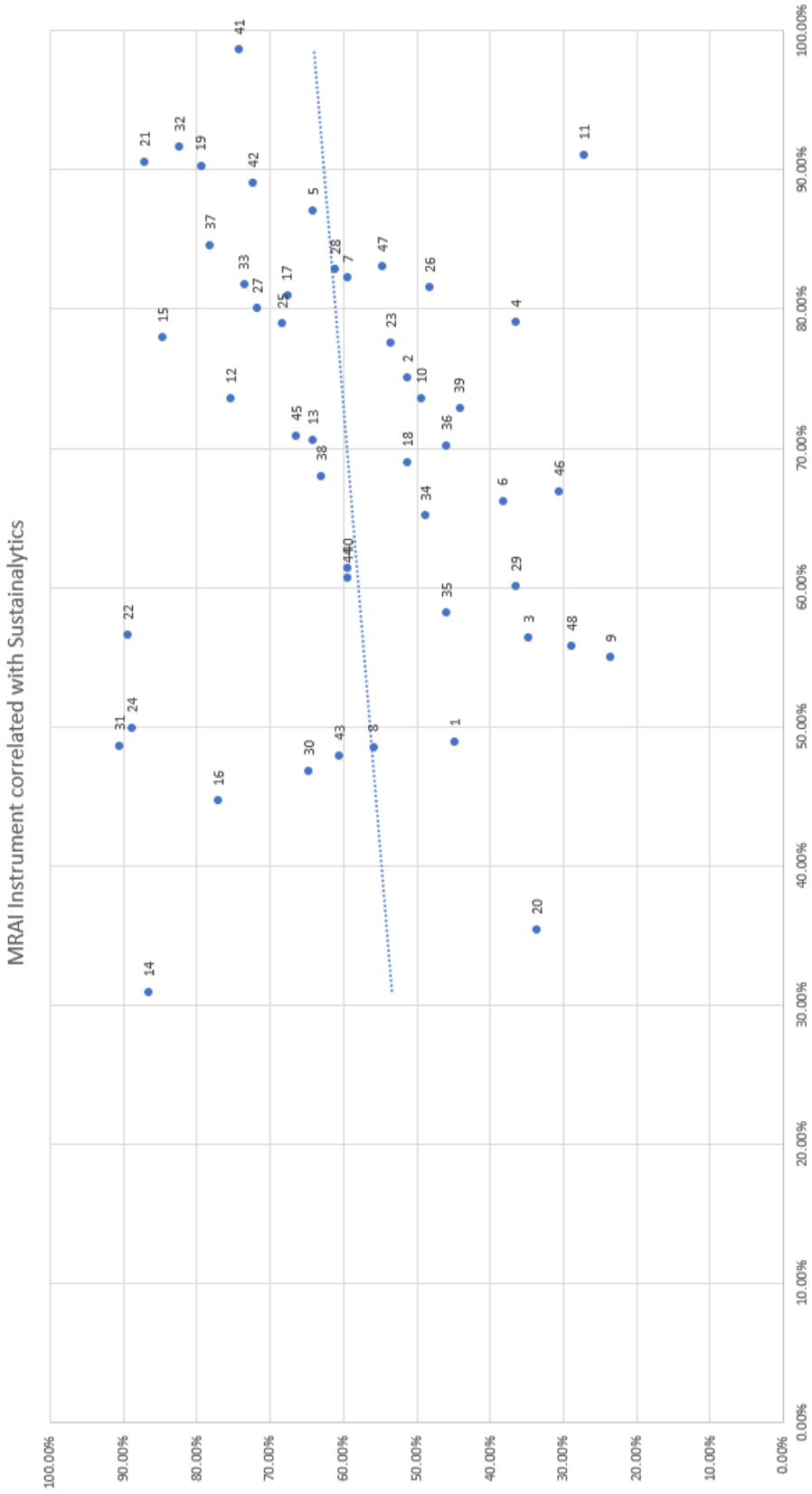


Figure 9: MRAI Instrument vs. Instrument ESG correlation

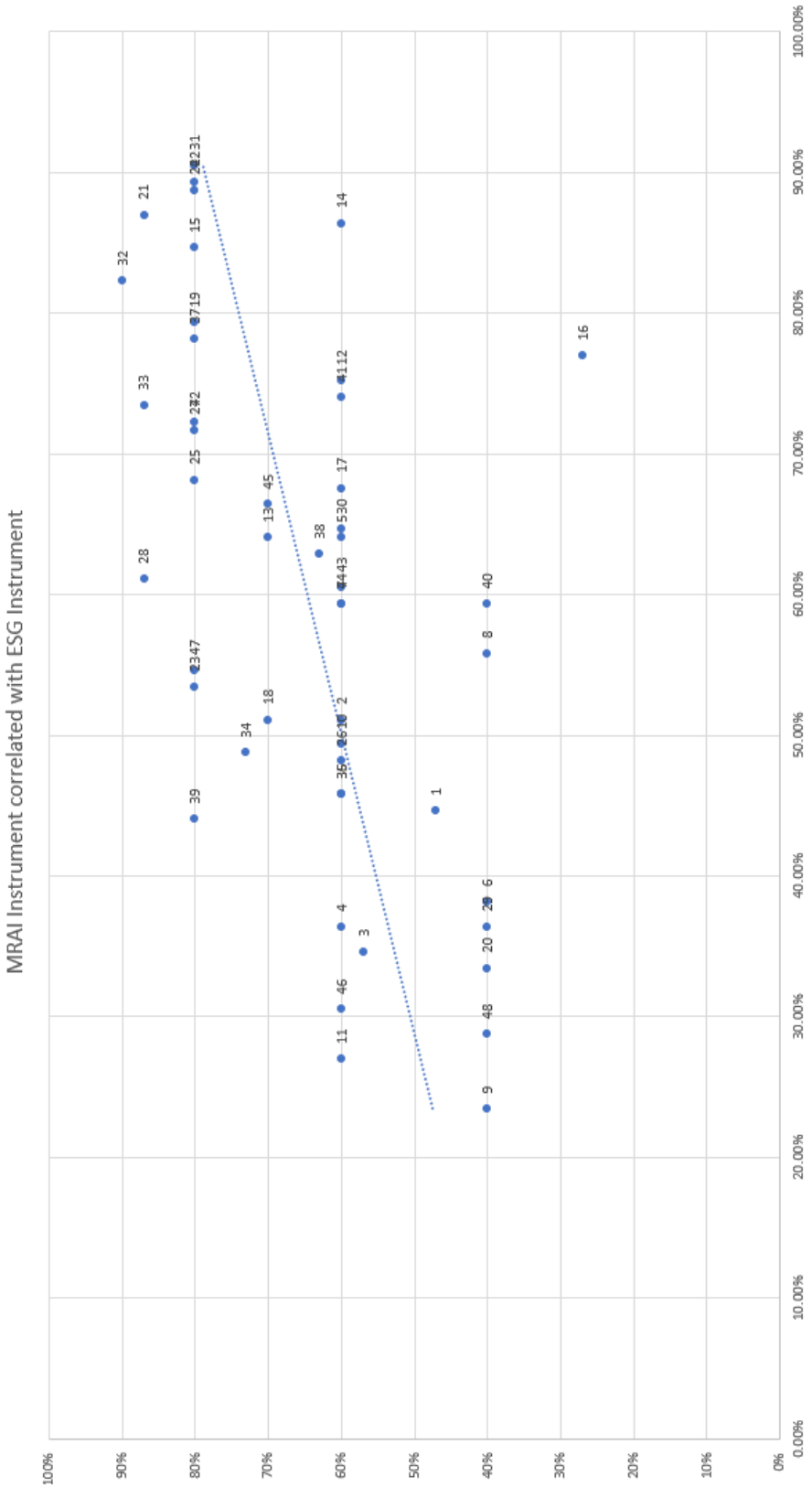


Table 9 - MTMM Regression Correlation Comparisons:

Below is a table summarizing the key comparisons in the different traits and methods demonstrating construct validity.

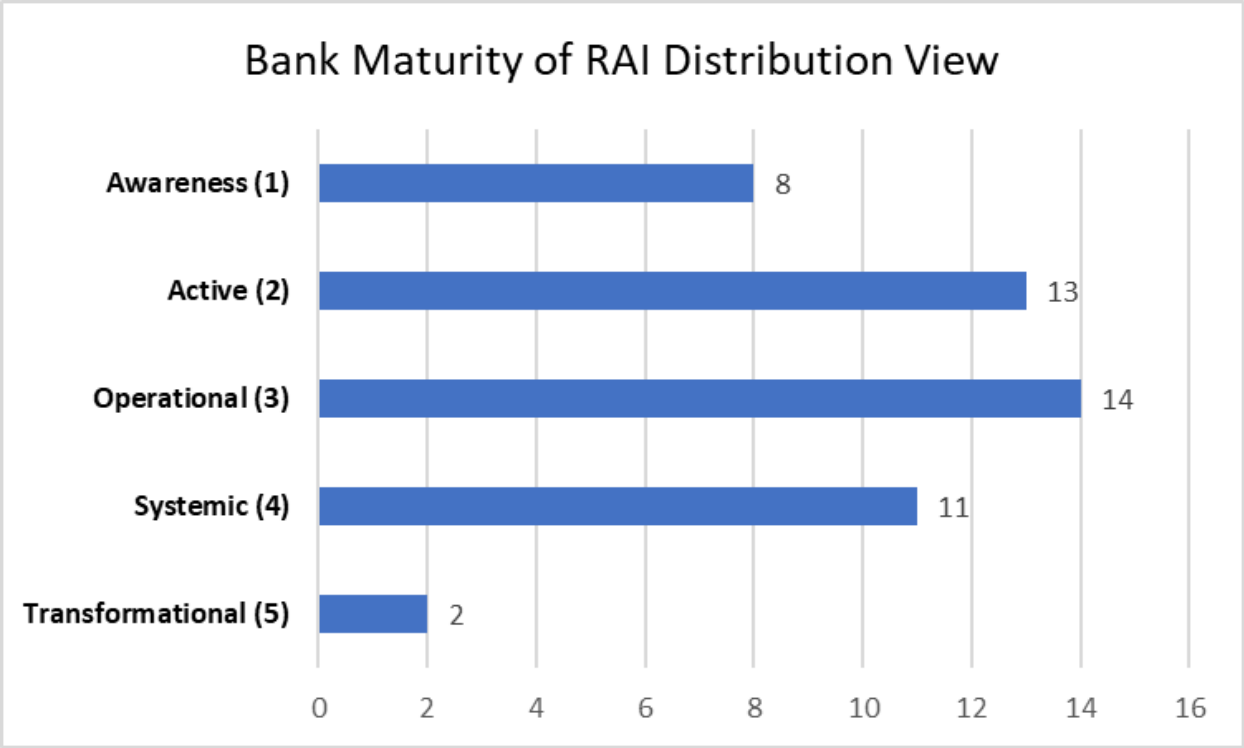
MTMM Regression Correlation Comparisons			
1	**Instrument MRAI-Proxy MRAI** mono-trait - multi-method (.882)	>	**Instrument MRAI-Instrument ESG** multi-trait - mono-method (.553)
2	**Instrument MRAI-Proxy MRAI** mono-trait - multi-method (.882)	>	**Proxy MRAI-Instrument ESG** multi-trait - multi-method (.398)
3	**Instrument MRAI-Proxy MRAI** mono-trait - multi-method (.882)	>	**Instrument MRAI-Sustainalytics ESG** multi-trait - multi-method (.135)
4	**Instrument MRAI-Proxy MRAI** mono-trait - multi-method (.882)	>	**Instrument ESG - Sustainalytics ESG** mono-trait - multi-method (.532)
5	**Instrument MRAI-Instrument ESG** multi-trait - mono-method (.553)	>	**Proxy MRAI-Instrument ESG** multi-trait - multi-method (.398)
6	**Instrument MRAI-Instrument ESG** multi-trait - mono-method (.553)	>	**Instrument ESG - Sustainalytics ESG** mono-trait - multi-method (.532)
7	**Instrument ESG - Sustainalytics ESG** mono-trait - multi-method (.532)	>	**Proxy MRAI-Instrument ESG** multi-trait - multi-method (.398)
8	**Instrument ESG - Sustainalytics ESG** mono-trait - multi-method (.532)	>	**Instrument MRAI-Sustainalytics ESG** multi-trait - multi-method (.135)

4.8 Capability Maturity Model

Building upon the original CMM (Paulk, Curtis, & Chrissis, 2001), however, leveraging the more recent Gartner AI Maturity Model (Gartner & Panetta, 2019) for the maturity category labels, this study calculated the average between the two MRAI data points (Instrument and Proxy), and then translated the score into a capability measure between 1-5. The labels in the graph correspond to the CMM rating (1 – Awareness, 2 – Active, 3 – Operational, 4 – Systemic, 5 – Transformational). The graph in **Figure 5** depicts the distribution of the Banks' maturity levels from this study's statistics according to the blend of

the MRAI (Instrument and Proxy) scores. The lower blended scores were associated with 1s and 2s and the higher blended scores were associated with 4s and 5s. The distribution skews to the lower side of maturity, which is consistent with the general sentiment during the interviews and is supported by **Figure 7** depicting the correlation between Instrument MRAI and Proxy MRAI reflecting that there is a range of maturity within the various Banks in 2022 as well as the mean of the MRAI scores which is 53.74%.

Figure 10: Bank MRAI instrument survey data in Gartner AI CMM category format.



4.9 Normal Distribution Curve of Average Maturity

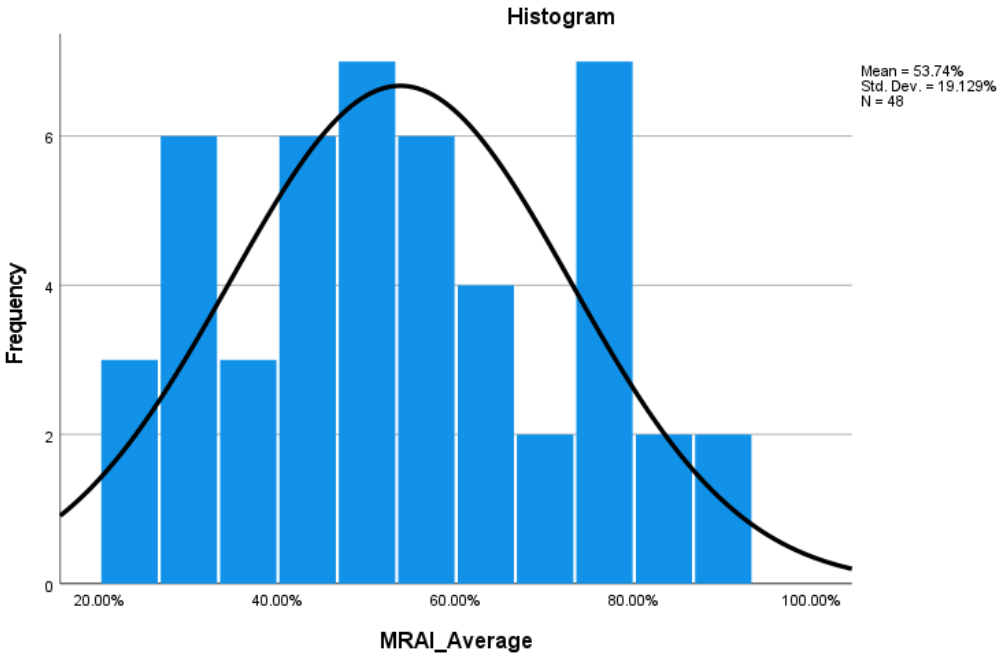
In addition to the CMM chart in **Figure 10**, this study ran an additional statistical analysis that demonstrates the distribution in **Figure 11** of the scores across the MRAI average, which is an average of the Instrument score and the Proxy score. With a mean of 53.74%, this is consistent with the CMM chart in which the majority of Banks are in the **"Operational"** maturity status, signaling, in general, that there is room for improvement across the Banks in their RAI maturity.

Figure 11: MRAI Instrument Statistical Distribution

Statistics

MRAI_Average

N	Valid	48
	Missing	0
Mean		53.7354%
Median		51.8750%
Mode		20.66%[a]
Std. Deviation		19.12882%
Variance		365.912
Skewness		.124
Std. Error of Skewness		.343
Range		71.51%



CHAPTER 5

DISCUSSION & CONCLUSIONS

5.1 Overview

Responsible AI has become a critical topic over the past decade in conjunction with a focus on responsible business (Lacy, Long, & Spindler, 2020), and a focus on DE&I (diversity, equity, and inclusion) and fairness driven by Bank's ESG (Environmental, Social, Governance) agendas (Gillan, Koch, & Starks, 2021). Since one of the main and most profitable revenue generators for Bank's is their credit lending (for example mortgages, auto loans, and credit cards) (Abedifar et al., 2018; Le & Ngo, 2020), there is a constant push for innovation and efficiency in these business processes. AI has been deployed in a significant capacity for Banks engaged in credit lending (Biswas et al., 2020), thus with the aforementioned focus on fairness in lending, responsible AI has become a central capability to focus on in credit lending. This study has addressed a gap in the industry by inventing a new measurement instrument (MRAI), with which Banks can now assess the maturity of their responsible AI program and capabilities.

5.2 Theoretical Contributions

This study engaged in a comprehensive research effort to canvas the most commonly referenced RAI principles (**Table 1**), and then the key researcher made judgements to include the most relevant principles as categories of the MRAI instrument. This new instrument addressed a gap in the industry regarding the ability perform a reliable and valid quantitative assessment on the maturity of a Banks' RAI capabilities. The study incorporated the following categories (Explainability, Fairness, Data Mgmt., and Security) based on the % ranking from the principle's analysis (**Figure 4**). Notably, the study added a novel category of 'Organizational Commitment', which incorporates elements of accountability, culture, strategy, investment, and decision-making. This category may prove to be the most important one for Banks in their quest to develop mature RAI capabilities (Ransbotham et al., 2019). The key contribution

of this paper is the introduction of a statistically validated new measurement instrument (MRAI) that Banks can deploy to assess their RAI maturity in credit lending capabilities. A second contribution is in addition to the assessment instrument, as part of the MTMM construct validity measurement, this study created another novel assessment tool named Proxy MRAI Instrument, which as previously described is a public reference review of some key indicators of leadership focus on RAI. Each of these instruments could be refined, generalized, and scaled to be applied to many more firms outside the Banking industry. This new MRAI instrument advances RAI research by enabling Banks to perform self-assessments to better prepare themselves for regulatory examinations and to be more proactive about explaining their fairness in lending in marketing communications.

These instruments are important for a couple of reasons. First, in the context of financial borrowers seeking an unbiased and fair credit lending process with a Bank, the Bank may use the MRAI score to include positive fairness messaging in their marketing materials. Secondly, in the context of the Algorithmic Accountability Act (Wyden et al., 2022), and other regulatory measures (Burt, 2021), Banks would be well advised to proactively build these explanatory capabilities to be able to nimbly respond to regulatory requests. Lastly, a third contribution is the general assessment of MRAI in the Banking industry across 48 of the top Banks in the US. This study has assessed the Banks with both MRAI instruments (Survey and Proxy) and created a statistical analysis of the overall maturity (**Figure 11**) as well as a regression correlation graph (**Figure 7**) that depicts the relationships and range in terms of the Bank's capabilities. In addition, this study applied a methodology based on the original CMM model (Paulk et al., 2001), and then advanced the CMM RAI **(Figure 10)** by leveraging a CMM model for AI developed by Gartner (Gartner & Panetta, 2019) **(Figure 6)** to focus specifically on levels of maturity relative to AI given the context of this study in assessing the maturity of RAI. In summary, the contributions from this research are two MRAI assessment instruments as well as a general assessment and CMM model of the current state of the Banking industry in the US based on the comprehensive analysis of 48 of the top Bank's RAI capabilities.

An important aspect of this study relates to the data used to perform the statistical analysis. The study proposal had aimed to interview 50 of the top US Banks (ADVRatings, 2021), however, two of the Banks were not able to be included in the data set due to lack of interview availability. The set of 48 of the top US Banks represent not just a data sample, but a nearly complete data population of the important Banks which creates a more robust analysis as these Banks comprise a significant majority of the industry in terms of customers, assets under management, and loans provided (ADVRatings, 2021). The CMM model **(Figure 10)** and the statistical distribution analysis captured the general disposition (**Figure 11**) of the Banking industry in that there was a mean of 53.74% maturity across the range of Banks, and the CMM model **(Figure 10)** had the most frequency in the "**Operational**" level of maturity, which was described by Gartner (Gartner & Panetta, 2019) as "*AI in production, creating value by e.g., processing optimization or product/service innovations*". The evidence provided through this lens of the data illustrates the state of maturity of RAI in the Banking industry today.

The statistical results from the data collection focused on three elements. First was the validation of the MRAI instrument categories and structure into itself, which was performed by computing Cronbach's alpha and CFA analysis on the elements of the instrument based on the initial interview with the 48 Banks' MRM executives. In terms of the instrument categories and structure, Cronbach's alpha was high at a value of .889 for the pre-validation instrument. The CFA analysis on the pre-validation instrument found one element (Weighting) that scored under .5, so this element was removed, and the CFA scores improved without this element. As a result, there was no special weighting applied to any of the categories in the MRAI instrument. Second was the validation of the attributes within the categories which was conducted by administering the MRAI instrument through survey interviews with the 48 Banks' MRM executives. Cronbach's alpha was computed for each of the categories with their respective attributes. Cronbach's alpha for the categories of Organizational Commitment was high at .898, and Explainability was high at .951, and Fairness was high at .931, and Data Mgmt. was high at .947, and lastly Security was high at .892. In terms of the CFA analysis for the MRAI instrument, the CFA for each

of the instrument items was above .6, with the lowest value (Data - CDO) being .638 and the highest value (Fairness – Data Training) being .873. Lastly, for the ESG questions survey, Cronbach's alpha was high at .877, and the CFA analysis demonstrated that all of the elements were above .5.

There was an additional element of robustness integrated into the study with the creation of the MRAI Proxy instrument. The significance of the MRAI Proxy is two-fold in that first, it is a separate instrument that could be enhanced with automation tools and scaled to a much wider set of companies by employing a text or semiotic analysis AI capability (https://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/). Secondly, the Cohen kappa result of .965 is extremely high in terms of agreement alignment. This result corresponds to agreement of 375 out of the 384 data elements. From a straight calculation perspective, this is 97.7% agreement, but then to account for the chance agreement, the Cohen kappa calculation provided a value of .965 or 96.5%. The reason for the Cohen kappa being high also has two components of rationale. First, a robust step by step validation script was provided by the key researcher to the raters for the inter-rater reliability test. Second, specific references to the evidence were provided to the raters such that they could click on a link from the reference recording data spreadsheet and review the requirements for the coding that was recorded for each of the raters.

Both the MRAI Proxy as well as the ESG questions survey served as data inputs into the MTMM analysis, which provided supporting evidence for convergent validity with a high correlation of .882 for the (mono-trait multi-method) test of Instrument MRAI and Proxy MRAI. This value was higher than the Instrument MRAI correlation with Sustainalytics (multi-trait multi-method) with a value of .135 providing supporting evidence for discriminant validity. There was another instance of discriminant validity (multi-trait mono-method) with the Instrument MRAI and Instrument ESG correlation of .553. Lastly, another validation with the (mono-trait multi-method) of Instrument ESG and Sustainalytics was higher at .532 than the (multi-trait multi-method) of Instrument ESG and Proxy MRAI at .398.

The above results demonstrate statistical evidence through three sets of Cronbach's alphas, and three sets of CFA analyses. These analyses coupled with the robustness of the Proxy MRAI inter-rater reliability along with the evidence of construct validity of the MTMM analysis provide for an extremely robust and validated MRAI instrument that Banks can use to assess the maturity of current RAI capabilities and create actions plans for both Marketing as well as Regulatory readiness.

5.3 Applied Implications

With the availability of the new MRAI instrument for Banks to leverage for assessing the maturity of the RAI capabilities they possess in credit lending, there is a significant opportunity to apply this instrument for benefit of the Bank. As has been discussed throughout this paper, AI is now ubiquitously pervasive in Bank credit lending processes as well as numerous other business process functions. Since the Banks aspire to continue to drive efficiency and productivity through the business processes (Ghosh et al., 2021), as well as improve accuracy of the algorithms they leverage, there will be continued investment into these capabilities (Borg, 2021). As AI continues to permeate every process, the importance of employing AI responsibly will be ever more important.

There are a few potential areas of implications for stakeholders that Bank's interact with that could benefit from leveraging the MRAI instrument. Similar to ESG and CSR (Gillan et al., 2021), RAI could weigh on how both institutional investors as well as individual investors choose their portfolio holdings. RAI could also impact how investment research analysts write about the stocks, which could also have an influencing effect on the prior point. The extant literature has tenuous linkages between ESG-CSR and RAI both from a function of how RAI is included in ESG analyses, as well as the connection of the principles between the RAI and ESG (Minkkinen, Niukkanen, & Mäntymäki, 2022). In terms of fairness and SRI (socially responsible investing) principles, there seems to be synergies with the conceptual nature of how issues can impact investments (Gadhoum, 2022; von Wallis & Klein, 2014). Investors could look up the MRAI score for a given company to determine if they deemed it worthy to invest in. In addition, the analysts who cover the stocks could leverage the MRAI score at face value or

even delve more deeply into the individual components of the score in the analyst reports on the stock, which, in turn creates an indirect accountability focus from the Banks on working toward a higher MRAI.

RAI will be so important that there will be increased regulatory governance over its usage (Burt, 2021; Candelon et al., 2021; Crosman, 2022; MacCarthy, 2020; Truby et al., 2020). This has implications for both executives at the Banks as well as policy makers that govern the Banks. A first implication for Banks of deploying the MRAI instrument will be for managers to perform self-assessments on various processes and capabilities that will be under regulatory scrutiny. This will address the S in ESG around the social implications of fairness in credit lending (de Laat, 2021). The hypothesis is that the more robust the RAI capabilities for a Bank are, the fairer that their lending processes have the potential to be. This will enable Banks to be better prepared to present processes and capabilities as well as respond to regulatory requests. The Algorithmic Accountability Act of 2022 (Wyden et al., 2022) is one of the key signals that the increased regulatory environment will be present in the coming years.

A second managerial implication again emphasizes the point around fairness in credit lending, but in this scenario, a Bank would leverage the MRAI assessment score to advertise through marketing communications that the Bank is socially conscious in line with ESG and CSR statements and provide a certified MRAI score in the marketing messages to advertise the fairness in lending. A third implication for executives is relevant as well, in that as stated above, with AI becoming ever more pervasive, leadership decisions will increasing be data driven through AI (Ransbotham et al., 2019; Stone et al., 2020). Responsible decision-making as a part of responsible business will be key for executives.

Regulators have been pushing for instrumentation such as the MRAI instrument that allow for the examiners to peer into opaque processes that Banks have in their lending practices. With the MRAI instrument certified and available, the regulators may be able to have a wider reach and more intimacy with the actual tensions that exist within the Bank. The policy makers and the executives could collaborate around the MRAI instrument and ensure there is absolute clarity about the credit lending decisions with a goal toward fair lending.

Lastly, the implication for the borrowers is significant, as with the industrialization of an MRAI instrument, the Banks can both comply with regulatory definitions with fairness in lending, as well as optimize the profitability for the Bank via the pareto efficiency frontier (Martinez et al., 2020) resulting in more chances given to those that typically may not have received the loans. Making decisions responsibly will be key to future leadership, and the ability to assess capabilities of organizational commitment, explainability, fairness, data quality, and security in the AI processes will be differentiating for future leadership.

5.4 Limitations and Future Research

This study was comprised of the development of a novel MRAI survey instrument as well as the usage of the MRAI instrument to assess maturity of Banks for their capabilities in Responsible AI. This study went through a comprehensive review and great lengths to review the principles it included in the instrument; however, it is certainly possible that other principles could be incorporated that would be more relevant or suitable in the view of some practitioners.

This study also performed multiple statistical analyses through Cronbach's alpha as well as CFA analyses to ensure the robustness and cohesiveness of the attributes in relation to the categories. In terms of the category and element validations as part of the pre-validation process, there could be a perceived limitation in the self-assessment survey method by which the instrument validation feedback was collected. This study did not challenge the scoring from the practitioners, and instead, collected the scores at face value. This same potential limitation is true of the MRAI instrument administration as the data was collected from the MRM executives as a self-assessment judgement. This data also served as a validation statistically that the categories and attributes were robust as components of the instrument.

In addition, a similar limitation exists in the Proxy MRAI instrument, as the study collected data from public websites into the eight categories the study chose. It is possible that there could have been different categories or more or less than eight categories to record the data. It is also very possible that some of the coding is incorrect based on the information that is available. That fact that the information is

not publicly available doesn't necessarily mean that the capabilities being assessed do not exist for a given Bank. Accordingly, with the world moving so quickly, and this data being collected in early 2022, there could be additional insights on principles, other instruments, RAI certification tools, as well as governmental policies that become available during the gestation of this paper being published.

The future research opportunities for this MRAI instrument are significant. First, this MRAI instrument could be slightly revised for various different business processes. While this study was focused on the credit lending process and some of the assessment attributes are very specific to these evaluations, the core categories and attributes could be applied to many different processes. As mentioned above in the applied implications, it is likely that data driven decision-making will be more prevalent leveraging AI and employing RAI. Second, this MRAI instrument score could become a standard independent variable for future research for which to be able to evaluate other aspects of companies, for example a correlation with ESG-CSR or Brand reputation indices. While this study superficially addressed this topic with involving the ESG variables in the MTMM analysis, there could be more extensive hypotheses and research in this area. Another idea in this genre is leveraging the MRAI score to correlate against financial performance in terms of P/E ratio or ROA. Third, there could be studies on how AI is impacting decision-making in terms of what capabilities to invest in at Banks and other companies expanding outside of Financial Services. For example, reviewing how Banks invest in Fraud detection AI capabilities as compared with how they invest in RAI capabilities.

Finally, as mentioned in the implications section, if responsibility becomes a key attribute of measurement for investing, partnering, or buying from responsible firms, then the MRAI score could become a fundamental tenet of business. With buyers, investors, and partners evaluating new ways of working with each-other based on criteria other than financial return, MRAI could be a new key indicator to a possible contagion effect of a higher standard for responsible business. In fact, a body of research around building responsibility into the design of key decision-making processes and capabilities (Benjamins et al., 2019) is already becoming popular with next generation firms.

CONCLUSION

As has been discussed throughout this study, AI is uniquely powerful, and being employed to derive myriad benefits throughout society. This study focused on RAI in Banks and addressed a gap in the academic and industry knowledge of the inability to measure RAI in the Banking industry by developing a new MRAI instrument. This new MRAI instrument will enable various stakeholders (investors, investment analysts, executives, policy makers, customers) in the Banking industry to measure RAI capabilities and build action plans for Banks based on their positioning within the CMM maturity curve spectrum.

This study ensured face and content validity of the new instrument through conducting a comprehensive study on the extant literature and building upon existing sets of RAI principles, frameworks, and toolkits. Based on the key principles, this study incorporated the significant categories as well as supplemented additional granular attributes to create and validate the MRAI instrument with interviews with Model Risk Managers within the Banks. This study further assured reliability and validity of the new instrument through robust statistical analysis on the categories and attributes using Cronbach's alpha and CFA techniques and then administered the new MRAI instrument through Bank MRM survey interviews.

The result of the data collection then comprised a mosaic of the Banking industry which is depicted through a correlation graph in **Figure 7** as well as a CMM model in **Figure 10** and represents a general capability set that has room for improvement at a mean maturity of 53.74%. As in any academic research, there are a few limitations that are explained above, as well as many ideas for additional research. Responsible business and fairness in lending will continue to fuel interest and investment into new processes and capabilities with AI, and thus mandate additional focus and need for more maturity and responsibility in AI.

REFERENCES

Abedifar, P., Molyneux, P., & Tarazi, A. (2018). Non-interest income and bank lending. *Journal of Banking & Finance, 87*, 411-426. doi:10.1016/j.jbankfin.2017.11.003

Adam, M., Wessel, M., & Benlian, A. (2020). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*. doi:10.1007/s12525-020-00414-7

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., . . . Venkatasubramanian, S. (2017). Auditing black-box models for indirect influence. *Knowledge and Information Systems, 54*(1), 95-122. doi:10.1007/s10115-017-1116-3

ADVRatings. (2021). Top 50 Banks in America. Retrieved from https://www.advratings.com/banking/top-banks-in-the-us

Aequitas. (2019). Bias and Fairness Audit Toolkit. Retrieved from http://aequitas.dssg.io/

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*: Harvard Business Review Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). An Economic Perspective on Artificial Intelligence: Unpacking the Challenges of Human-Machine Interaction. *NATO Defense College*, 11.

AIEthicist. (2021). AI Frameworks, Guidelines, Toolkits. *AI Frameworks.* Retrieved from https://www.aiethicist.org/frameworks-guidelines-toolkits

Al-Rubaie, M., & Chang, J. M. (2019). Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security & Privacy, 17*(2), 49-58. doi:10.1109/msec.2018.2888775

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a Convolutional Neural Network. *ICET2017*.

Almeida, P. G. R. d., dos Santos, C. D., & Farias, J. S. (2021). Artificial Intelligence Regulation: a framework for governance. *Ethics and Information Technology, 23*(3), 505-525. doi:10.1007/s10676-021-09593-z

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*.

Alves, G., Amblard, M., Bernier, F., Couceiro, M., & Napoli, A. (2021). *Reducing Unintended Bias of ML Models on Tabular and Textual Data*. Paper presented at the 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA).

Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). Customer experiences in the age of artificial intelligence. *Comput Human Behav, 114*, 106548. doi:10.1016/j.chb.2020.106548

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989. doi:10.1177/1461444816676645

Anderson, M., & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine, 28*(4), 12.

Anthi, E., Williams, L., Rhode, M., Burnap, P., & Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. *Journal of Information Security and Applications, 58*. doi:10.1016/j.jisa.2020.102717

Arnold, T., & Scheutz, M. (2018). The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology, 20*(1), 59-69. doi:10.1007/s10676-018-9447-7

Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. *IBM Research AI, 1*(10).

Asilomar, Institute, F. o. L., & FLI. (2017). Asilomar - AI Intelligent Machines Smart Policies. *Future of Life*. Retrieved from futureoflife.org/ai-principles

Askell, A., Brundage, M., & Hadfield, G. (2019). The Role of Cooperation in AI Development. *arXiv:1907.04534v1*.

Ayling, J., & Chapman, A. (2021). Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*. doi:10.1007/s43681-021-00084-x

Baktha, K., & Tripathy, B. K. (2017). Investigation of Recurrent Neural Networks in the field of sentiment analysis. *International Conference on Communication and Signal Processing*.
Balasubramanian, N., Ye, Y., & Xu, M. (2019). Substituting Human Decision-Making with Machine Learning: Implications for Organizational Learning. *Academy of Management Review, 70*(4), 46.
Bamberger, P. A. (2018). AMD—Clarifying What We Are about and Where We Are Going. *Academy of Management Discoveries, 4*(1), 1-10. doi:10.5465/amd.2018.0003
Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405-417. doi:10.1016/j.eswa.2017.04.006
Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. doi:10.1016/j.inffus.2019.12.012
Barros, R. S. M. d., & Santos, S. G. T. d. C. (2019). An overview and comprehensive comparison of ensembles for concept drift. *Information Fusion, 52*, 213-244. doi:10.1016/j.inffus.2019.03.006
Barth, J. R., Lin, C., Ma, Y., Seade, J., & Song, F. M. (2013). Do bank regulation, supervision and monitoring enhance or impede bank efficiency? *Journal of Banking & Finance, 37*(8), 2879-2892. doi:10.1016/j.jbankfin.2013.04.030
Baum, J. A. C., & Haveman, H. A. (2020). Editors' Comments: The Future of Organizational Theory. *Academy of Management Review, 45*(2), 268-272. doi:10.5465/amr.2020.0030
Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943v1*.
Benjamins, R. (2020). A choices framework for the responsible use of AI. *AI and Ethics, 1*(1), 49-53. doi:10.1007/s43681-020-00012-5
Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by Design in Practice. *AAAI Proceedings - Telefonica - arXiv:1909.12838*.
Biswas, S., Carson, B., Chung, V., Singh, S., & Thomas, R. (2020). AI-bank of the future: Can banks meet the AI challenge? *McKinsey&Company, 1*(2020), 14.
Boddington, P. (2017). *Toward a Code of Ethics for Artificial Intelligence*.
Boddington, P., Millican, P., & Wooldridge, M. (2017). Minds and Machines Special Issue: Ethics and Artificial Intelligence. *Minds and Machines, 27*(4), 569-574. doi:10.1007/s11023-017-9449-y
Boden, M. A. (2016). *AI: Its nature and future* (Vol. 1). UK: Oxford University Press.
Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection- A Review. *Institute of Mathematical Statistics, 17*(3), 235-249.
Borg, J. S. (2021). Four investment areas for ethical AI: Transdisciplinary opportunities to close the publication-to-practice gap. *Big Data & Society*.
Bostrom, N. (2005). A History Of Transhumanist Thought. *Journal of Evolution and Technology, 14*, 25.
Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*: Oxford University Press, Inc.
Boza, P., & Evgeniou, T. (2021). Implementing Ai Principles- Frameworks, Processes, and Tools. *Insead Working Paper*.
Broniatowski, D. A. (2021). Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. *NIST*. doi:10.6028/nist.Ir.8367
Brown, E., & Piroska, D. (2021). Governing Fintech and Fintech as Governance: The Regulatory Sandbox, Riskwashing, and Disruptive Social Classification. *New Political Economy, 27*(1), 19-32. doi:10.1080/13563467.2021.1910645
Brynjolfsson, E., & McAfee, A. (2011). *Race Against the Machine*. Lexington, MA: Digital Frontier Press.
Brynjolfsson, E., & McAfee, A. (2012). Winning the Race With Even Smarter Machines. *MITSloan Management Review, 1*(12), 1-23.
Brynjolfsson, E., & McAfee, A. (2016). *The Second Machine Age: Work, Progress and Prosperity In a Time of Brilliant Technologies*. New York: W. W. Norton & Company.

Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. *National Bureau of Economic Research, 24*(1), 46.
Buchanan, B. G. (2006). A (Very) Brief History of Artificial Intelligence. *AI Magazine, 26*(4).
Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society, 64*. doi:10.1016/j.techsoc.2020.101475
Bülbül, D., Hakenes, H., & Lambert, C. (2019). What influences banks' choice of credit risk management practices? Theory and evidence. *Journal of Financial Stability, 40*, 1-14. doi:10.1016/j.jfs.2018.11.002
Burgt, J. v. d. (2019). Explainable AI in Banking. *Journal of Digital Banking, 4*(4), 344-350.
Burkhardt, R., Hohn, N., & Wigley, C. (2019). Leading your organization to responsible AI. *McKinsey Analytics*.
Burt, A. (2021). New AI Regulations Are Coming. Is Your Organization Ready? Retrieved from https://hbr.org/2021/04/new-ai-regulations-are-coming-is-your-organization-ready
Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable Machine Learning in Credit Risk Management. *Computational Economics, 57*(1), 203-216. doi:10.1007/s10614-020-10042-0
Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & Finance, 72*, 218-239. doi:10.1016/j.jbankfin.2016.07.015
Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the MultiTrait-Multimethod Matrix. *Psychological Bulletin*.
Campbell, M. (2019). Synthetic Data: How AI Is Transitioning From Data Consumer to Data Producer... and Why That's Important. *Computer, 52*(10), 89-91. doi:10.1109/mc.2019.2930097
Campbell, M., Hoane, J. J., & Hsu, F.-h. (2002). Deep Blue. *Artificial Intelligence, 134*(1-2), 57-83.
Candelon, F., Carlo, R. C. d., Bondt, M. D., & Evgeniou, T. (2021). AI Regulation Is Coming. Retrieved from https://hbr.org/2021/09/ai-regulation-is-coming
Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos Trans R. Soc. A, 376*(2133). doi:10.1098/rsta.2018.0080
Cavello, B. (2020). PAI Launches Interactive Project To Put Ethical AI Principles into Practice. *Partnership for AI (PAI)*. Retrieved from https://partnershiponai.org/pai-launches-interactive-project-to-put-ethical-ai-principles-into-practice/
Certo, S. T., Lester, R. H., Dalton, C. M., & Dalton, D. R. (2006). Top Management Teams, Strategy and Financial Performance: A Meta-Analytic Examination. *Journal of Management Studies, 43*(4), 813 - 838.
CFPB. (2011). CFPB Consumer Laws and Regulations. *CFPB.gov*. Retrieved from https://files.consumerfinance.gov/f/documents/102012_cfpb_unfair-deceptive-abusive-acts-practices-udaaps_procedures.pdf
Chander, A., Srinivasan, R., Chelian, S., Wang, J., & Uchino, K. (2018). Working with Beliefs: AI Transparency in the Enterprise. *Fujitsu*.
Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36*(4), 1165-1188.
Chen, J. (2018). *Fair lending needs explainable models for responsible recommendation.* Paper presented at the Proceedings of Workshop on Responsible Recommendation, Vancouver, Canada.
Cheng, L., Varshney, K. R., & Liu, H. (2021). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. *arXiv:2101.02032v3*.
Chowdhury, R., & Williams, J. (2021). Introducing Twitter's first algorithmic bias bounty challenge. Retrieved from https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge
Christian, B., & Griffiths, T. (2016). *Algorithms to Live By: The Computer Science of Human Decisions*: Henry Holt and Co., Inc.

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. *Information, 12*(7). doi:10.3390/info12070275
Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law & Security Review, 35*(4), 410-422. doi:10.1016/j.clsr.2019.04.007
Coates, D. L., & Martin, A. (2019). An instrument to evaluate the maturity of bias governance capability in artificial intelligence projects. *IBM Journal of Research and Development, 63*(4/5), 7:1-7:15. doi:10.1147/jrd.2019.2915062
Coeckelbergh, M. (2020). *AI Ethics*. Cambridge, MA: MIT Press.
Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.
Congress.gov. (2019). A bill to direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments. *Congress.gov.* Retrieved from https://www.congress.gov/bill/116th-congress/senate-bill/1108/text
Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness - A critical review of Fair Machine Learning. *ArVix - 1808.00023*.
core-econ, & Liebniz. (2021). The Pareto Efficiency Curve. Retrieved from https://www.core-econ.org/the-economy/book/text/leibniz-05-08-01.html
Cowgill, B. (2019). Bias and Productivity in Humans and Machines. *Columbia - Working Paper*.
Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine, 35*(1), 53-65. doi:10.1109/msp.2017.2765202
Crosman, P. (2022). Will regulators' warnings chill lenders' use of AI? Retrieved from https://www.americanbanker.com/news/will-cfpb-federal-reserve-warnings-chill-lenders-use-of-ai
Daugherty, P. R., & Wilson, H. J. (2018). *Human+Machine*. Boston, MA: Harvard Business Review Press.
Daugherty, P. R., & Wilson, H. J. (2022). *Radically Human*. Boston, MA: Harvard Business Review Press.
Davenport, T., & Faccioli, G. (2019). Cognitive on the Continent. Retrieved from https://www.tomdavenport.com/cognitive-on-the-continent/
Davenport, T., & Kirby, J. (2016). Just How Smart Are Smart Machines. *MITSloan Management Review, 1*(21), 1-17.
Davenport, T. H., & Ronanki, R. (2018). Artificial Intelligence For The Real World: Don't Start With Moon Shots. *Harvard Business Review*, 1-10.
Davis, J. L., Williams, A., & Yang, M. W. (2021). Algorithmic reparation. *Big Data & Society, 8*(2). doi:10.1177/20539517211044808
de Castro Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. *Applied Soft Computing, 83*. doi:10.1016/j.asoc.2019.105640
de Laat, P. B. (2021). Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? *Philos Technol*, 1-59. doi:10.1007/s13347-021-00474-3
Deepa, B., & Ramesh, K. (2021). *Production Level Data Pipeline Environment for Machine Learning Models*. Paper presented at the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS).
Dell. (2020). cdo perspectives how to achieve data management maturity. *Dell Technologies CDO Perspectives*.
Demšar, J., & Bosnić, Z. (2018). Detecting concept drift in data streams using model explanation. *Expert Systems with Applications, 92*, 546-559. doi:10.1016/j.eswa.2017.10.003

Deng, Y., & Gabriel, S. (2006). Risk-Based Pricing and the Enhancement of Mortgage Credit Availability among Underserved and Higher Credit-Risk Populations. *Journal of Money, Credit and Banking, 38*(6), 1431-1460.
Deon. (2021). An ethics checklist for data scientists. Retrieved from https://deon.drivendata.org/
Digital-Solutions. (2021). ROI of AI: The Cost-Benefit of your next Project. Retrieved from https://daitan.com/blog-post/roi-of-ai-the-cost-benefit-of-your-next-project/
Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*.
Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). *Measuring and Mitigating Unintended Bias in Text Classification*. Paper presented at the Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.
Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *XIV*.
Dowling, G., & Moran, P. (2012). Corporate Reputation: Built in or Bolted on? *CALIFORNIA MANAGEMENT REVIEW, 54*(2), 25-42.
Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63-71. doi:10.1016/j.ijinfomgt.2019.01.021
Durmus, M. (2021). A brief Overview of some Ethical-AI Toolkits. *Medium.* Retrieved from https://medium.com/nerd-for-tech/an-brief-overview-of-some-ethical-ai-toolkits-712afe9f3b3a
Dwork, C., Rothblum, G. N., & Vadhan, S. (2010). *Boosting and Differential Privacy*. Paper presented at the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science.
Economist. (2020). Staying Ahead Of The Curve: The business case for Responsible Artificial Intelligence. *Economist Intelligence Unit, 1*.
Eitel-Porter, R. (2020). Beyond the promise: implementing ethical AI. *AI and Ethics, 1*(1), 73-80. doi:10.1007/s43681-020-00011-6
Element.AI. (2021). The AI Maturity Framework: A strategic guide to operationalize and scale enterprise AI solutions. *ElementAI.* Retrieved from https://www.elementai.com/
Emmert-Streib, F., Yli-Harja, O., & Dehmer, M. (2020). Explainable artificial intelligence and machine learning: A reality rooted perspective. *10*(6), e1368. doi:https://doi.org/10.1002/widm.1368
Etzioni, A., & Etzioni, O. (2017). Should Artificial Intelligence Be Regulated? *Issues in Science and Technology, 33*(4), 32-36.
EUCommission. (2019). The Assessment List for Trustworthy Artificial Intelligence. *HLEG*.
Fatemi, A., & Fooladi, I. (2006). Credit Risk Management: a survey of practices. *Managerial Finance, 32*(3).
Fay, B. (2021). Fair Credit Reporting Act. *Debt.org*.
FederalReserve.gov. (2011a). SR 11-7 Guidance on Model Risk Management. *Board of Governors of the Federal Reserve System, 1*, 1 - 21.
FederalReserve.gov. (2011b). Supervisory Guidance on Model Risk Management SR 11-7. Retrieved from https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf
Feng, X., Jiang, Y., Yang, X., Du, M., & Li, X. (2019). Computer vision algorithms and hardware implementations: A survey. *Integration, 69*, 309-320. doi:10.1016/j.vlsi.2019.07.005
Ferràs-Hernández, X. (2017). The Future of Management in a World of Electronic Brains. *Journal of Management Inquiry, 27*(2), 260-263. doi:10.1177/1056492617724973
Fifth-Quadrant. (2021). The Responsible AI Index. Retrieved from https://www.fifthquadrant.com.au/2021-responsible-ai-index
Figini, S., Bonelli, F., & Giovannini, E. (2017). Solvency prediction for small and medium enterprises in banking. *Decision Support Systems, 102*, 91-97. doi:10.1016/j.dss.2017.08.001
Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Berkman Klein Center for Internet & Society*.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Retrieved from https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/7
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach (Dordr), 28*(4), 689-707. doi:10.1007/s11023-018-9482-5
Follett, J. (2019). How 22 Years of AI Superiority Changed Chess. *Toward Data Science*.
Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-Powered Organization: Technology isn't the biggest challenge. Culture is. *Harvard Business Review, 7*(1), 1-13.
Främling, K., Westberg, M., Jullum, M., Madhikermi, M., & Malhi, A. (2021). Comparison of Contextual Importance and Utility with LIME and Shapley Values. In *Explainable and Transparent AI and Multi-Agent Systems* (pp. 39-54).
Friedman, G., & McCarthy, T. (2020). Employment Law Red Flags in the Use of Artificial Intelligence in Hiring. Retrieved from https://www.americanbar.org/groups/business_law/publications/blt/2020/10/ai-in-hiring/
Friedman, M. (1970). A Friedman Doctrine - The Social Responsibility Of Business Is To Increase Its Profits. *NY Times*, 3.
ftc.gov. (2020a). Credit Reporting. Retrieved from https://www.ftc.gov/news-events/media-resources/consumer-finance/credit-reporting
ftc.gov. (2020b). Using Artificial Intelligence and Algorithms. Retrieved from https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms
ftc.gov. (2021). Aiming for truth, fairness, and equity in your company's use of AI. Retrieved from https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai
Gadhoum, Y. (2022). Artificial Intelligence Trends and Ethics: Issues and Alternatives for Investors. *Intelligent Control and Automation, 13*(01), 1-15. doi:10.4236/ica.2022.131001
Garbuio, M., & Lin, N. (2018). Artificial Intelligence as a Growth Engine for Health Care Startups: Emerging Business Models. *CALIFORNIA MANAGEMENT REVIEW, 61*(2), 59-83. doi:10.1177/0008125618811931
Gartner, & Panetta, K. (2019). The CIO's guide to Artificial Intelligence. Retrieved from https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence
Ghadiri, D. P., Gond, J.-P., & Brès, L. (2015). Identity work of corporate social responsibility consultants: Managing discursively the tensions between profit and social responsibility. *Discourse & Communication, 9*(6), 32.
Ghosh, B., Prasad, R., & Pallail, G. (2021). *The Automation Advantage: Embrace the Future of Productivity and Improve Speed, Quality, and Customer Experience Through AI*.
Gillan, S. L., Koch, A., & Starks, L. T. (2021). Firms and social responsibility: A review of ESG and CSR research in corporate finance. *Journal of Corporate Finance, 66*. doi:10.1016/j.jcorpfin.2021.101889
GitHub. (2022). Where the world builds software. Retrieved from https://github.com/
Golbin, I., Rao, A. S., Hadjarian, A., & Krittman, D. (2020). *Responsible AI: A Primer for the Legal Community*. Paper presented at the 2020 IEEE International Conference on Big Data (Big Data).
Goleman, D. (1995). *Emotional intelligence / Daniel Goleman*. New York: Bantam Books.
Goo, J. J., & Heo, J.-Y. (2020). The Impact of the Regulatory Sandbox on the Fintech Industry, with a Discussion on the Relation between Regulatory Sandboxes and Open Innovation. *Journal of Open Innovation: Technology, Market, and Complexity, 6*(2). doi:10.3390/joitmc6020043
GooglePAIR. (2021). What-If Tool - Playing with AI Fairness. Retrieved from https://pair-code.github.io/what-if-tool/ai-fairness.html
Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk. *Front Artif Intell, 4*, 752558. doi:10.3389/frai.2021.752558
Gupta, A., Bhatt, D., & Pandey, A. (2021). *Transitioning from Real to Synthetic Data: Quantifying the Bias in the Model*. Paper presented at the Synthetic Data Generation Workshop - ICLR 2021.

Hafen, R., & Critchlow, T. (2013). *EDA and ML -- A Perfect Pair for Large-Scale Data Analysis*. Paper presented at the 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum.
Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*(1), 99-120. doi:10.1007/s11023-020-09517-8
Hall, P., Cox, B., Dickerson, S., Ravi Kannan, A., Kulkarni, R., & Schmidt, N. (2021). A United States Fair Lending Perspective on Machine Learning. *Front Artif Intell, 4*, 695301. doi:10.3389/frai.2021.695301
Harari, Y. N. (2017). Reboot for the AI revolution. *Nature.* Retrieved from https://www.nature.com/news/reboot-for-the-ai-revolution-1.22826
Hategan, C.-D., Sirghi, N., Curea-Pitorac, R.-I., & Hategan, V.-P. (2018). Doing Well or Doing Good: The Relationship between Corporate Social Responsibility and Profit in Romanian Companies. *Sustainability, 10*(4), 1041. doi:10.3390/su10041041
Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). *Improving Fairness in Machine Learning Systems*. Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.
Holzinger, A., Kieseberg, P., Weippl, E., & Tjoa, A. M. (2018). Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. *11015*, 1-8. doi:10.1007/978-3-319-99740-7_1
Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Retrieved from https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter
Hunter, A. P., Sheppard, L. R., Karlen, R., & Balieiro, L. (2018). Managing Operational Artificial Intelligence. *Center for Strategic and International Studies (CSIS)*.
Hunter, A. P., Sheppard, L. R., Karlen, R., & Baliero, L. (2018). Adoption of Artificial Intelligence. *Carnegie Endowment for International Peace, 2019*.
IBMAI. (2021). AI Explanabililty 360. Retrieved from http://aix360.mybluemix.net/
India, S. (2019). Google vs. Amazon vs. Microsoft vs. Facebook — Who is leading the AI race? *Medium.* Retrieved from https://medium.com/@springboard_ind/google-vs-amazon-vs-microsoft-vs-facebook-who-is-leading-the-ai-race-9e9cefbc5c45
Institute, R.-R. A. (2022). RAI Certification. Retrieved from https://www.responsible.ai/certification
Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018, 20-24 May 2018). *Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning.* Paper presented at the 2018 IEEE Symposium on Security and Privacy (SP).
Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577-586. doi:10.1016/j.bushor.2018.03.007
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399. doi:10.1038/s42256-019-0088-2
Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.
Kanaan, M. (2020). *T-Minus AI: Humanity's Countdown to Artificial Intelligence and the New Pursuit of Global Power*: BenBella Books.
Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25. doi:10.1016/j.bushor.2018.08.004
Kavanagh, C. (2019). Artificial Intelligence: New Tech, New Threats, and New Governance Challenges. *Carnegie Endowment for International Peace, 1*.
Kaya, O. (2019). Artificial Intelligence in Banking. *EU Monitor, 6*(4), 9.
Kearns, M., & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*: Oxford University Press, Inc.

Kelly, K. (2012). Better Than Human: Why Robots Will — And Must — Take Our Jobs. Retrieved from https://www.wired.com/2012/12/ff-robots-will-take-our-jobs/
Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767-2787. doi:10.1016/j.jbankfin.2010.06.001
Klein, A. (2021). Reducing bias in AI-based financial services. *Brookings AIET*. Retrieved from https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/
Koshiyama, A., Kazim, E., & Treleaven, P. (2021). Towards Algorithm Auditing A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. *SSRN*.
Kurzweil, R. (1999). *Age of Spiritual Machines: When Computers Exceed Human Intelligence*: Penguin USA.
Kurzweil, R. (2006). *The Singularity Is Near: When Humans Transcend Biology*: Penguin (Non-Classics).
Lacy, P., Long, J., & Spindler, W. (2020). *The Circular Economy Handbook: Realizing the Circular Advantage*: Palgrave Macmillan.
Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.
Langenbucher, K. (2020). Responsible AI Credit Scoring - A Legal Framework. *Fordham Law School*.
Le, T. D. Q., & Ngo, T. (2020). The determinants of bank profitability: A cross-country analysis. *Central Bank Review, 20*(2), 65-73. doi:10.1016/j.cbrev.2020.04.001
Leavitt, K., Schabram, K., Hariharan, P., & Barnes, C. (2019). Ghost in the Machine - On Organizational Theory in the Age of Machine Learning. *Academy of Management Review*, 72.
Lee, M., & Floridi, L. (2020). Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. *Minds and Machines*. doi:10.1007/s11023-020-09529-4
Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks, 7*(1). doi:10.3390/risks7010029
Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute*. doi:10.5281/zenodo.3240529
Liebergen, B. v. (2021). Machine Learning: A Revolution in Risk Management and Compliance? *Institute of International Finance*.
Liker, J. K., Haddad, C. J., & Karlin, J. (1999). Perspectives on Technology and Work Organization. *Annual Review of Sociology, 25*, 575 - 596.
Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy (Basel), 23*(1). doi:10.3390/e23010018
Lorica, B. (2018). Managing risk in machine learning. *O'Reilly*. Retrieved from https://www.oreilly.com/radar/managing-risk-in-machine-learning/
Lorica, B., Doddi, H., & Talby, D. (2019a). Managing machine learning in the enterprise: Lessions from banking and health care. *O'Reilly*. Retrieved from https://www.oreilly.com/radar/managing-machine-learning-in-the-enterprise-lessons-from-banking-and-health-care/
Lorica, B., Doddi, H., & Talby, D. (2019b). What are model governance and model operations? *O'Reilly*. Retrieved from https://www.oreilly.com/radar/what-are-model-governance-and-model-operations/
Loukides, M. (2016). To supervise or not to supervise in AI? Retrieved from https://www.oreilly.com/radar/to-supervise-or-not-to-supervise-in-ai/
Lovelock, J.-D. (2020). Gartner Says Worldwide IT Spending to Grow 4% in 2021. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2020-10-20-gartner-says-worldwide-it-spending-to-grow-4-percent-in-2021
Lu, N., Zhang, G., & Lu, J. (2014). Concept drift detection via competence models. *Artificial Intelligence, 209*, 11-28. doi:10.1016/j.artint.2014.01.001

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst*, *30*, 4765–4774.
MacCarthy, M. (2020). AI needs more regulation, not less. Retrieved from https://www.brookings.edu/research/ai-needs-more-regulation-not-less/
Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research, 120*, 262-273. doi:10.1016/j.jbusres.2020.07.045
Martinez, N. (2021). Blind Pareto Fairness and Subgroup Robustness. *No Journal*.
Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax Pareto Fairness- A Multi Objective Perspective. *Proc Mach Learn Res, 119*, 6755–6764.
Maurer, R. (2020). AI-Based Hiring Concerns Academics, Regulators. Retrieved from https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/ai-based-hiring-concerns-academics-regulators.aspx
Mayer-Schnberger, V. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think. Viktor Mayer-Schnberger and Kenneth Cukier*: John Murray Publishers.
McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine, 27*(4).
MetaAI. (2021). How we're using Fairness Flow to help build AI that works better for everyone. Retrieved from https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/
MicrosoftFairlearn. (2021). Fairness in Machine Learning. Retrieved from https://fairlearn.org/main/user_guide/mitigation.html
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38. doi:10.1016/j.artint.2018.07.007
Mills, S., & Duranton, S. (2021). Are You Overestimating Your Responsible AI Maturity? Retrieved from https://www.bcg.com/publications/2021/the-four-stages-of-responsible-ai-maturity
Minevich, M. (2020). 4 Ways That You Can Prove ROI From AI. Retrieved from https://www.forbes.com/sites/markminevich/2020/03/03/4-ways-that-you-can-prove-roi-from-ai/?sh=84de94e784a7
Minkkinen, M., Niukkanen, A., & Mäntymäki, M. (2022). What about investors? ESG analyses as tools for ethics-based AI auditing. *Ai & Society*. doi:10.1007/s00146-022-01415-0
Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the IRE, 49*(1).
Misheva, B. H., Hirsa, A., Osterrieder, J., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in credit risk management. *arXiv:2103.00949v1*.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., . . . Gebru, T. (2019). Model Cards for Model Reporting. *arXiv:1810.03993v2*. doi:10.1145/3287560
Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2020). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *arXiv:1811.07867v3*.
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501-507. doi:10.1038/s42256-019-0114-4
Moore, G. E. (1965). Cramming more components onto integtrated circuits. *Electronics, 38*(8), 33-35.
Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Boston, MA: Harvard University Press.
Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *https://arxiv.org/abs/1905.06876*.
Murawski, J. (2019). Mortgage Providers Look to AI to Process Home Loans Faster. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/mortgage-providers-look-to-ai-to-processhome-loans-faster-11552899212

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A, 116*(44), 22071-22080. doi:10.1073/pnas.1900654116

Murphy, J. W., & Largacha-Martínez, C. (2021). Is it possible to create a responsible AI technology to be used and understood within workplaces and unblocked CEOs' mindsets? *Ai & Society*. doi:10.1007/s00146-021-01316-8

Myers, G., & Nejkov, K. (2020). Developing Artificial Intelligence Sustainably: Toward a Practical Code of Conduct for Disruptive Technologies. *EM Compass, 80*.

Newborn, M., & Newborn, M. (1997). *Kasparov Vs. Deep Blue: Computer Chess Comes of Age*: Springer-Verlag.

NSTC. (2016). The National Artificial Intelligence Research and Development Strategic Plan. *www.whitehouse.gov/ostp/nstc*.

O'Neil, C. (2016). *Weapons of Math Destruction*. New York: Broadway Books.

OPENAI. (2021). OPEN AI WebSite. Retrieved from https://openai.com/

Pacelli, V., & Azzollini, M. (2011). An Artificial Neural Network Approach for Credit Risk Management. *Journal of Intelligent Learning Systems and Applications, 03*(02), 103-112. doi:10.4236/jilsa.2011.32012

Page, J., Bain, M., & Mukhlish, F. (2018, 24-27 Aug. 2018). *The Risks of Low Level Narrow Artificial Intelligence.* Paper presented at the 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR).

Pan, Y. (2016). Heading toward Artificial Intelligence 2.0. *Engineering, 2*(4), 409-413. doi:10.1016/j.Eng.2016.04.018

Papernot, N., & Brain, G. (2018). A Marauder's Map of security and privacy in Machine Learning. *arXiv:1811.01134v1, 1*.

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2016). SoK: Towards the Science of Security and Privacy in Machine Learning. *arXiv:1611.03814v1, 1*.

Pasiouras, F., Tanna, S., & Zopounidis, C. (2009). The impact of banking regulations on banks' cost and profit efficiency: Cross-country evidence. *International Review of Financial Analysis, 18*(5), 294-302. doi:https://doi.org/10.1016/j.irfa.2009.07.003

Pattberg, P. (2006). The Influence of Global Business Regulation: Beyond Good Corporate Conduct. *Business and Society Review, 111*(3), 241–268.

Paulk, M., Curtis, B., & Chrissis, M. B. (2001). *Capability Maturity Model for Software*.

Paulk, M. C., Curtis, B., Chrissis, M. B., & Weber, C. V. (1993). Capability Maturity Model for Software, Version 1.1. *Technical Report CMU/SEI-93-TR-024*.

Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data Lifecycle Challenges in Production Machine Learning: A Survey. *SIGMOD Record, 47*(2), 12.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., . . . Iyengar, S. S. (2019). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys, 51*(5), 1-36. doi:10.1145/3234150

Prince, A. E. R., & Schwarcz, D. (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review, 105*(3).

Rahman, N., & Blake, L. (2021). A review of CSR classification schemes and the operationalization of bolted-on vs. built-in CSR. *Business Ethics, the Environment & Responsibility, 30*(3), 248-261. doi:10.1111/beer.12345

Rahwan, I. (2017). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*(1), 5-14. doi:10.1007/s10676-017-9430-8

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., . . . Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477-486. doi:10.1038/s41586-019-1138-y

Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137-141. doi:10.1007/s11747-019-00710-5

Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation-Augmentation Paradox. *Academy of Management Review, 48*(1), 192-210.
Raj, S. B. E., & Portia, A. A. (2011, 18-19 March 2011). *Analysis on credit card fraud detection methods.* Paper presented at the 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET).
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., . . . Barnes, P. (2020). *Closing the AI accountability gap*. Paper presented at the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.
Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2020). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Association for Computing Machinery, 1*(1), 23. doi:10.1145/nnnnnnn.nnnnnnn
Ransbotham, S., Khodabandeh, S., Fehling, R., LaFountain, B., & Kiron, D. (2019). Winning With AI. *MIT Sloan Review*.
Rao, A., & Golbin, I. (2019). What is fair when it comes to AI bias? Retrieved from strategy-business.com/media/file/What-is-fair-when-it-comes-to-AI-bias.pdf
Rao, S. S. (1987). Game theory approach for multiobjective structural optimization. *Computers & Structures, 25*(1), 119-127. doi:https://doi.org/10.1016/0045-7949(87)90223-9
Renda, A. (2019). Artificial Intelligence Ethics, Governance and Policy Challenges. *CEPS*.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?"*. Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
Rist, L. (2018). Encrypt your Machine Learning. *Medium.* Retrieved from https://medium.com/corti-ai/encrypt-your-machine-learning-12b113c879d6
Robertson, J., Diab, A., Marin, E., Nunes, E., Paliath, V., Shakarian, J., & Shakarian, P. (2016). Darknet Mining and Game Theory for Enhanced Cyber Threat Intelligence. *The Cyber Defense Review, 1*(2), 95-122.
Rodriguez, L. (2020). All Data is Not Credit Data. *Columbia Law Review, 120*(7), 1843 - 1884.
Rodriguez, M., de Araújo, L. J. P., & Mazzara, M. (2020). Good practices for the adoption of DataOps in the software industry. *Journal of Physics: Conference Series, 1694*. doi:10.1088/1742-6596/1694/1/012032
Rossi, F. (2019). Building Trust in Artificial Intelligence. *Journal of International Affairs, 72*(1), 127 - 134.
Roszbach, K. (2003). Bank lending policy, credit scoring, and the survival of loans. *The Review of Economics and Statistics, 86*(4), 946-958.
Russell, S. J., Hauert, S., Altman, R., & Veloso, M. (2015). Ethics of Artificial Intelligence. *Nature, 521*(7553), 415-416.
Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
Russell, S. J., Norvig, P., & Intelligence, A. (2009). *Artificial Intelligence: A Modern Approach 3rd Edition*. Englewood Cliffs, NJ: Prentice Hall.
Sadiq, R. B., Safie, N., Abd Rahman, A. H., & Goudarzi, S. (2021). Artificial intelligence maturity model: a systematic literature review. *PeerJ Comput Sci, 7*, e661. doi:10.7717/peerj-cs.661
Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., . . . Ghani, R. (2019). Aequitas - A Bias and Fairness Audit Toolkit. *arXiv:1811.05577v2*.
Saleiro, P., Stevens, A., Anisfeld, A., & Ghani, R. (2018). The Fairness Tree. Retrieved from http://www.datasciencepublicpolicy.org/projects/aequitas/
Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 5-22). Cham: Springer International Publishing.
Sandmo, A. (1970). Equilibrium and Efficiency in Loan Markets. *Economica, 37*(145), 23-38. doi:10.2307/2551999

Satell, G., & Abdel-Magied, Y. (2020). AI Fairness Isn't Just an Ethical Issue. *Harvard Business Review*(October 2020).
Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2020). Principles to Practices: Closing the responsible AI gap. *arXiv:2006.04707v1*.
Sen, P., & Ganguly, D. (2020). Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(03), 2685-2692. doi:10.1609/aaai.v34i03.5654
Shestakofsky, B. (2017). Working Algorithms: Software Automation and the Future of Work. *Work and Occupations, 44*(4), 376-423. doi:10.1177/0730888417726119
Shrestha, A., & Mahmood, A. (2019). Review of Deep Learning Algorithms and Architectures. *IEEE Access, 7*, 53040-53065. doi:10.1109/access.2019.2912200
Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics. *Journal of Database Management, 31*(2), 74-87. doi:10.4018/jdm.2020040105
Silver, D., & Tesauro, G. (2009). *Monte-Carlo simulation balancing*. Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada. https://doi.org/10.1145/1553374.1553495
Simon, H. A. (1947). *Administrative Behavior*. New York: The MacMillan Company.
Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics, 69*(1), 99-118.
Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4*(3), 181-201.
Simon, H. A. (1987). Making management decisions: The role of intuition and emotion. *The Academy of Management Executive, 1*(1), 57-68.
Singh, P. J., Franceschini, F., & Smith, A. (2006). An empirically validated quality management measurement instrument. *Benchmarking: An International Journal, 13*(4), 493-522. doi:10.1108/14635770610676317
Spangler, W. E. (1991). The Role of Artificial Intelligence in Understanding the Strategic Decision-Making Process. *IEEE Transactions on Knowledge and Data Engineering, 3*(2), 149 - 160.
Spy, A. (2018). *The Workplace of the Future*. Retrieved from: https://www.economist.com/leaders/2018/03/28/the-workplace-of-the-future
Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., . . . Machtynger, L. (2020). Artificial intelligence (AI) in strategic marketing decision-making: a research agenda. *The Bottom Line, 33*(2), 183-200. doi:10.1108/bl-03-2020-0022
Stoyanovich, J., Howe, B., & Jagadish, H. V. (2020). Responsible data management. *Proceedings of the VLDB Endowment, 13*(12), 3474-3488. doi:10.14778/3415478.3415570
Taddeo, M. (2017). Trusting Digital Technologies Correctly. *Minds and Machines, 27*(4), 565-568. doi:10.1007/s11023-017-9450-5
Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *CALIFORNIA MANAGEMENT REVIEW, 61*(4), 15-42. doi:10.1177/0008125619867910
Tammenga, A. (2020). The application of Artificial Intelligence in banks in the context of the three lines of defence model. *Maandblad Voor Accountancy en Bedrijfseconomie, 94*(5/6), 219-230. doi:10.5117/mab.94.47158
Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY: Penguin (Vintage) Books.
Teichmann, J. (2019). Bias and Algorithmic Fairness. Retrieved from https://towardsdatascience.com/bias-and-algorithmic-fairness-10f0805edc2b
Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence, 302*. doi:10.1016/j.artint.2021.103627
Tjong Tjin Tai, T. F. E. (2016). The right to be forgotten – private law enforcement. *International Review of Law, Computers & Technology, 30*(1-2), 76-83. doi:10.1080/13600869.2016.1138628

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society, 63*. doi:10.1016/j.techsoc.2020.101413
Truby, J., Brown, R., & Dahdal, A. (2020). Banking on AI: mandating a proactive approach to AI regulation in the financial sector. *Law and Financial Markets Review, 14*(2), 110-120. doi:10.1080/17521440.2020.1760454
Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind - A Quarterly Review of Psychology and Philosphy, 59*(236), 433-460.
Vakkuri, V., Jantunen, M., Halme, E., Kemell, K.-K., Nguyen-Duc, A., Mikkonen, T., & Abrahammson, P. (2021). *Time for AI Ethics Maturity Model is Now*. Paper presented at the Workshop for Artificial Intelligence Safety (SafeAI 2021).
Vakkuri, V., Kemell, K.-K., & Kultanen, J. (2020). The Current State of Industrial Practice in Artificial Intelligence Ethics. *IEEE SOFTWARE*.
Vanian, J. (2021). Why synthetic data is such a hot topic in the artificial intelligence world. *Fortune, 12-22-21*.
Vogel, D. (2008). Private Global Business Regulation. *Annual Review of Political Science, 11*(1), 261-282. doi:10.1146/annurev.polisci.11.053106.141706
Vohra, A.-S. (2022). The art of AI maturity. Retrieved from https://www.accenture.com/us-en/insights/artificial-intelligence/ai-maturity-and-transformation
von Krogh, G. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries, 4*(4), 404-409. doi:10.5465/amd.2018.0084
von Wallis, M., & Klein, C. (2014). Ethical requirement and financial interest: a literature review on socially responsible investing. *Business Research, 8*(1), 61-98. doi:10.1007/s40685-014-0015-7
Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening The Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 842 - 861.
Wall, L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business, 100*, 55-63. doi:https://doi.org/10.1016/j.jeconbus.2018.05.003
Wang, X., Li, J., Kuang, X., Tan, Y.-a., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing, 130*, 12-23. doi:10.1016/j.jpdc.2019.03.003
Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning. *Procedia Computer Science, 174*, 141–149.
Weber, M. (1905). *The Protestant ethic and the spirit of capitalism / Max Weber ; new introduction and translation by Stephen Kalberg*. Chicago ; London: Fitzroy Dearborn.
Wee, C. K., & Nayak, R. (2019). A novel machine learning approach for database exploitation detection and privilege control. *Journal of Information and Telecommunication, 3*(3), 308-325. doi:10.1080/24751839.2019.1570454
Welch, J. (2001). *Jack : straight from the gut / Jack Welch ; with John A. Byrne*. New York: Warner Books.
Whang, S. E., Tae, K. H., Roh, Y., & Heo, G. (2020). Responsible AI Challenges in End-to-end Machine Learning. *Need to Confirm - Confirm - Confirm*.
Wolf, J. (2020). How to improve cybersecurity for artificial intelligence. *Brookings AIET*. Retrieved from https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/
Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons, 61*(6), 823-832. doi:10.1016/j.bushor.2018.07.001
Wyden, Booker, & Clarke. (2022). The Algorithmic Accountability Act of 2022. Retrieved from https://www.wyden.senate.gov/download/one-pager-bill-summary-of-the-algorithmic-accountability-act-of-2022

Xu, Y., Shieh, C.-H., van Esch, P., & Ling, I. L. (2021). AI Customer Service: Task Complexity, Problem-Solving Ability, and Usage Intention. *Australasian Marketing Journal, 28*(4), 189-199. doi:10.1016/j.ausmj.2020.03.005

Zackova, E. (2015). Intelligence Explosion Quest for Humankind. *Beyond artificial intelligence: The disappearing human-machine divide, 9*, 31-43. doi:10.1007/978-3-319-09668-1_3

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun, 10*(1), 3770. doi:10.1038/s41467-019-11786-6

Zaib, M., Sheng, Q. Z., & Emma Zhang, W. (2020). *A Short Survey of Pre-trained Language Models for Conversational AI-A New Age in NLP*. Paper presented at the Proceedings of the Australasian Computer Science Week Multiconference.

Zeng, Y., Lu, E., & Huanfu, C. (2019). Linking Artificial Intelligence Principles. *AAAI Proceedings on Artificial Intelligence Safety*.

Zetzsche, D. A., Arner, D., Buckley, R., & Tang, B. W. (2020). Artificial Intelligence in Finance - Putting the Human in the Loop. *CFTE (Center for Finance, Technology and Entrepreneurship) Academic Paper Series, 1*, 50.

Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., Chen, J., & Sudjianto, A. (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. *Wells Fargo*.

Zhu, L., Qiu, D., Ergua, D., Yinga, C., & Liu, K. (2019). *A study on predicting loan default based on the random forest algorithm.* Paper presented at the 7th International Conference on Information Technology and Quantitative Management.

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*(4), 1060-1089. doi:10.1007/s10618-017-0506-1