AI and Ethics
https://doi.org/10.1007/s43681-023-00321-5

ORIGINAL RESEARCH

Measuring responsible artificial intelligence (RAI) in banking: a valid and reliable instrument

John Ratzan[1] · Noushi Rahman[1]

Received: 4 April 2023 / Accepted: 5 July 2023
© The Author(s) 2023

Abstract
Widespread use of artificial intelligence (AI) and machine learning (ML) in the US banking industry raises red flags with regulators and social groups due to potential risk of data-driven algorithmic bias in credit lending decisions. The absence of a valid and reliable measure of responsible AI (RAI) has stunted the growth of organizational research on RAI (i.e., the organizational balancing act to optimize efficiency and equity). To address this void, we develop a novel measurement instrument to assess RAI maturity in firms. A review of the nascent literature reveals that there is a wide distribution of RAI capabilities. The RAI instrument that we advance is based on the exhaustive review of this dispersed literature. Analyses of data from large US banks show strong evidence of validity and reliability of the RAI maturity instrument.

Keywords Artificial intelligence (AI) · Responsible AI · Machine learning (ML) · AI ethics · Bias · Fairness

1 Introduction

Artificial intelligence (AI) has been ubiquitously adopted by Fortune 500 companies in their quest to leverage big data insights to optimize various aspects of their businesses [53, 76, 127]. In parallel, competition has increased organizational pressure to create competitive advantage from AI initiatives in terms of speed, efficiency, effectiveness, expense optimization, profitability, and ROI [13, 17, 44]. Scholars have shown that AI can be effective for corporations in myriad ways, for example, in predicting mortgage loan payments [109], fighting fraud [16], preventing adversarial security breaches [108], optimizing employee hiring [59], and automating virtual agent customer service [2].

We define RAI as the ability to implement AI and ML [40, 79] models that transparently explain data inputs and predicted outputs while maintaining fairness (i.e., mitigating bias and harm). RAI focuses on ensuring the ethical, transparent, and accountable use of AI technologies in a manner consistent with user-expectations, organizational values, societal laws, and norms [41]. RAI is also often comprised of a set of principles for organizations, such as fairness, explainability, privacy, and security which are present in the key literature [43, 54, 62, 95, 115]. Fairness defined in this context as equity for the individual stakeholders engaged in the given interaction is the foundation for RAI in that it aims to remove bias from the AI decision process, which follows the definition for bias from the Equal Credit Opportunity Act (ECOA) (i.e. prohibiting discrimination in any aspect of a credit transaction) [57]. As defined, fairness and bias seem to have a negatively correlated relationship in that when bias rises, fairness declines [72]. Explainability and transparency provide a platform for evaluating fairness because if the AI process is explainable, then organizations are more apt to address bias removal [24, 74, 104]. Other key principles concern data privacy and data security encompassed by data management and data quality, which are paramount to the integrity of the AI process [80, 97]. There are also RAI principles that are related to management, accountability, and governance, however, an overarching and cohesive principle to measure the significance of these items is missing. We introduce a new category named 'organizational commitment' encompassing accountability, governance, financial investment, leadership, diversity, humanity, culture, employee engagement, and training.

We neither discount accountability nor consider governance RAI principles as unimportant. While accountability

✉ John Ratzan
john.ratzan@gmail.com

Noushi Rahman
nrahman@pace.edu

[1] Pace University—Lubin School of Business, New York, USA

Published online: 11 September 2023

and governance are necessary RAI principles and present in nearly all the related literature, we argue that those principles are not sufficient to ensure that leadership commitment, culture, and financial investments are present to properly deploy RAI [89]. The rationale for including accountability and governance in the organizational commitment principle is to enable the instrument to measure a succinct list of key principles, while ensuring that there is a leadership focus on RAI. Organizational commitment encompasses and supports accountability by incorporating a culture of RAI and awareness training programs in mature RAI deployments [28]. The 'organizational commitment' principle may be the most critical component to focus on in developing RAI capabilities [21, 106].

RAI is critical for managing the complex web of compliance with ESG (environmental, social, and governance) obligations and business regulations [22, 78, 113]. While endeavoring to achieve competitive advantage and associated ROI [103], organizations must possess mature RAI programs that enable them to balance the tension between optimizing AI for accuracy and supporting fairness to serve an equitable, social purpose [12, 111].

With AI still in an initial adoption phase, RAI is only partially implemented in various organizational processes, affording limited opportunity to explain, interpret, and understand the nature of RAI [64, 65]. Due to the nascent nature of AI capability development, there are only scant public references on investment costs for implementing AI governance programs which can range from simple auditing [93] to fully mature RAI programs [3, 15, 66, 128]. However, the increasing rate of AI adoption has bolstered governance pressure to ensure that AI/ML has a core set of principles incorporated into organizational values [21, 38].

Most of the extant literature focuses on the impact of more narrow forms of AI (e.g., credit underwriting, automated customer service virtual agents, and employee hiring processes) or on productivity and profit [5, 9, 23, 116]. Scholars have noted various methods to assess RAI maturity [8, 11, 125]. For the purposes of our research, RAI maturity is represented by possessing the capabilities to address fairness and transparency in various organizational processes. Although a few works review the inherent biases [96] or describe RAI frameworks [125] or discuss RAII (Responsible AI Institute) certifications, these works do not deal with RAI measurement. The absence of an instrument to measure firm-level maturity of RAI is a critical gap in the pertinent academic literature. Such an instrument will accelerate empirical research on RAI and aide firms to uniformly implement RAI. We address this gap in the literature by developing an instrument to assess RAI maturity and argue that firms that exhibit robust RAI maturity should score highly in our RAI instrument.

We reviewed the RAI programs and capabilities that banks possess to manage the powerful risks and benefits of AI in credit lending through the development of a survey instrument to measure RAI maturity. We contribute to the literature in four ways. First, we introduce a novel instrument with which to measure RAI maturity by building on extant RAI principles literature. Second, within the instrument, in addition to incorporating extant commonly accepted principles from the literature, we introduce a new RAI principle named 'organizational commitment'. Third, we created a supplementary proxy instrument leveraging archival data to satisfy the construct and convergent validity focus of our research. Lastly, we analyze RAI and ESG, and conceptually separate these related, but different measures in support of providing evidence for discriminant validity.

2 A prefatory note on the banking industry context

When employing the power of AI, corporations risk introducing biases into automated decision-making, which attracts scrutiny of regulatory agencies [27, 112]. Regulators are concerned with the Fair Credit Reporting Act (FCRA) and the Equal Credit Opportunity Act (ECOA) for fair lending. Moreover, the Supervisory Guidance on Model Risk Management (SR11-7) defines Federal Reserve (Fed) governance as the regulation inherent in technology model usage in fair lending. This regulatory oversight is similar in nature to the 2018 GDPR (General Data Protection Regulation) present in the EU (European Union). Ethical bias in data and algorithms can pervade the technical capabilities in various ways, including errors, oversights, or unintended consequences with the concern that AI deployed on a large scale can adversely impact certain groups and individuals [82, 96]. There are different types of data to consider for fairness in credit decisioning. For example, there are data related to credit bureau information and application data about the potential borrower as well as alternate data that generally possesses additional attributes that may be leveraged by the creditor [67, 84]. While there are risks of potential bias in both types of data, the larger risk exists in alternate data [61]. This rapid advancement and adoption of AI is taxing the capacity of banks to leverage the emerging technology while ensuring compliance with regulatory measures [47, 73].

Banks earn a significant amount of their profit through credit lending in mortgage underwriting, providing auto loans, and issuing credit cards [1]. Each of the credit lending underwriting decisions effectively assesses the risk that the individual borrowing the money will not re-pay the loan [92, 123]. 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 [73].

Decision makers involved with the credit lending process rely on the AI analysis tools to balance profitability with fairness in the credit lending decisions [13]. AI technology is computationally efficient in traditional logistical regression-based credit systems and also enables banks to perform advanced analytics on each potential borrower at near instantaneous speed. One of the key data elements used in analysis is the credit score managed by FICO (Fair Isaac Co.) [71], which has raised the alarm of fair lending bias risk with borrowers as well as activist groups [30].

3 Responsible AI (RAI)

RAI is about being responsible for the power of AI [41]. Coeckelbergh [33] conceptualizes RAI as AI Ethics. RAI is acutely relevant when human decision-making is delegated to AI [127]. RAI 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 [21, 81].

One of the key aspects of governing AI or determining whether AI is "responsible" is understanding the transparency and interpretability of the algorithm and model defined as explainable AI (XAI) [74, 104]. The design must engender trust and provide explainable transparency for the results from the data, model, and algorithm [120]. By providing clarity into the governance of AI components, RAI allows organizations to innovate responsibly to realize the transformative potential of AI [105].

Another key element of RAI is the care in the data lifecycle (data selection, data collection, and data management) such that bias does not creep into the overall system [98]. RAI is most critically needed in cases where there is a potential for bias or errors in the data, models, programs (algorithms), data training, and ensuring the proper governance controls are in place [31]. AI results should be examined for bias and provide evidence of fairness [113]. This includes cases where bias exists, but firms can offer some counterfactual explanatory evidence to support the AI decision [91].

Lastly, monitoring of the solution deployment is important to ensure that the results being generated match the intended design and adhere to the established governance [31, 63]. There are a few key elements involved in deploying AI in the organization such as abilities to maintain control, demonstrate fairness, ensure responsibility, and practice accountability for the capabilities [14, 41]. As part of the ongoing maintenance of the RAI programs, there are some components that need to be incorporated into an effective RAI program, such as performance drift monitoring [77], operational bias review [113] and model training [3].

Our review of RAI is built upon key research of RAI principles [18, 29, 38, 89, 115]. In terms of works which aggregate RAI principles, Jobin et al. [62] list 84 sources and summarize RAI principles into 11 categories comprising the most comprehensive and authoritative summary available. Another extensive summary was published in a matrix form [54] which listed 20 RAI principles. A work that contains a similar matrix summarizing both principles (listing 12 RAI principles) as well as fairness toolkits was published by the IFC EM Compass [95]. Lastly, another work published 8 RAI principles' categories and provided detailed rationale for each which is a technique that we leverage in our instrument explanation [43].

4 A measurement instrument for RAI

The survey instrument for measuring the maturity of RAI capabilities is presented in Table 1 of which contains the RAI maturity instrument in panel A and a brief measure of ESG for the purposes of subsequent validity testing in panel B. The instrument we aim to advance is structurally similar to a bias governance questionnaire instrument [32], and follows a known instrument development methodology [117]. The detailed rationale and key references are listed in Table 2.

We performed an analysis on the most common RAI principles which served as the foundation for the initial categorization of the instrument. This resulted in the five categories (i.e., organizational commitment, transparency, fairness, data management, and security) for the RAI instrument. Three of the categories (transparency, fairness, security) align directly with Jobin et al. [62], Hagendorff [54], Myers and Nejkov [95] and Fjeld et al. [43]. The two other categories (organizational commitment and data management) partially aligned to existing principles. We incorporate the 'governance and accountability' principles into a new principle called 'organizational commitment', as we argue that companies with strong support of RAI from the executive leadership will have more mature RAI programs [106]. Components such as leadership focus, financial investments, accountability, and culture are inherent in the organizational commitment category. In addition, financial investment, employee training, diversity priority, humanity focus, and financial ROI are central to governance and accountability within organizational commitment. RAI must include diverse participation to ensure that the AI systems will meet their societal and ethical principles [41]. Measuring these attributes in the 'organizational commitment' category creates a novel contribution to the RAI principles standard and differentiates our instrument from other RAI assessment

Table 1 RAI instrument structure

Panel A: RAI Maturity Instrument							
Factor Name	Key Evidence	Scoring					
Organizational Commitment to RAI	**This attribute measures the organizational commitment to RAI for accountability, governance, finance, culture**						
Organizational Structure RAI Focus	To what degree is there a formal org structure entity called Responsible or Ethical AI?	1	2	3	4	5	
Investment in RAI	To what degree is there evidence of significance financial investment linked to Responsible AI	1	2	3	4	5	
ROI Analysis on RAI	To what degree is there a formal financial ROI analysis performed on Responsible AI?	1	2	3	4	5	
Training for RAI	To what degree are there training programs in place for all employees on Responsible AI?	1	2	3	4	5	
Culture of AI Accountability	To what degree is there a perception of a culture of AI within the company?	1	2	3	4	5	
Pareto Efficiency Frontier Decisions	To what degree is pareto efficiency frontier analysis used in credit lending decisions?	1	2	3	4	5	
C-Suite Involvement	To what degree is the CEO or Board updated on the company's RAI program?	1	2	3	4	5	
Transparency & Explainability	**This attribute measures the degree of transparency in terms of the algorithms, and models that comprise the AI & ML**						
Explainability Governance	To what degree are there formal policies or processes in place to govern explainability?	1	2	3	4	5	
Regulatory Sandbox (Visibility)	To what degree are there capabilities in place to provide visibility to regulators on explainability?	1	2	3	4	5	
Model Audit Controls	To what degree are there capabilities in place to audit models?	1	2	3	4	5	
Model Drift Prevention Monitoring	To what degree are there capabilities or processes in place to test & mitigate model drift?	1	2	3	4	5	
Use of Knowledge Graphs	To what degree are knowledge graphs leveraged to provide model explainability?	1	2	3	4	5	
Use of Model Card Reporting	To what degree are model cards leveraged to provide a descriptive model explainability?	1	2	3	4	5	
Advanced ML Explain (LIME, SHAP)	To what degree is there use of advanced black box technology such as SHAP or LIME?	1	2	3	4	5	
Fairness/Bias Mitigation	**This attribute measures the ability to mitigate bias in lending and prevent discrimination harm**						
Policy for Fairness in Models	To what degree is there a governance policy to define the fairness rules in the models?	1	2	3	4	5	
Fairness in Training Data	To what degree are there fairness and governance considerations in place in the training data?	1	2	3	4	5	
Human in the Loop	To what degree are there HITL (Human in the Loop) governance in the ML workflow?	1	2	3	4	5	
Legal Implications	To what degree are you aware that there may be legal/compliance implications?	1	2	3	4	5	
Proxy Discrimination	To what degree does the capability to mitigate proxy discrimination exist?	1	2	3	4	5	
Model Reparation	To what extent are there processes to cure the bias if it is indeed found in the models?	1	2	3	4	5	
Data Management & Quality	**This attribute measures the maturity of the data mgmt processes that feeds the ML models**						
Data Privacy	To what degree are there data privacy considerations to protect PII?	1	2	3	4	5	
Differential Privacy Capability	To what degree are there differential privacy capability to protect PII in place?	1	2	3	4	5	
EDA for Pre-modeling	To what degree do exploratory data analysis processes exist as part of pre-modelling?	1	2	3	4	5	
Use of Data Pipeline Tools	To what degree are there data pipeline tools in use for the ML?	1	2	3	4	5	
Use of Big Data Lake	To what degree is there a modernized big data lake environment in place?	1	2	3	4	5	

Table 1 (continued)

Panel A: RAI Maturity Instrument							
Factor Name	Key Evidence	Scoring					
CDO Involvement	To what degree is the CDO (Chief Data Officer) intimately involved with model risk management?	1	2	3	4	5	
Use of Synthetic Data	To what degree does the capability exist to supplement data with synthetic data?	1	2	3	4	5	
Use of DataOps	To what degree is there a DataOps process in place for collaborative data management?	1	2	3	4	5	
Right to be Forgotten	To what degree is there a process in place to delete data for those who wish to be forgotten??	1	2	3	4	5	
Security	**This attribute measures what specific data privacy and data security provisions are in place?**						
Adversarial Cyber Attack Defense	To what degree are there ML adversarial attack defenses in place?	1	2	3	4	5	
Data Encryption	To what degree are there data encryption provisions in place to secure the data?	1	2	3	4	5	
Special Security Access for Production	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 Usage	To what degree are there security processes in place to prevent unintended use of AI?	1	2	3	4	5	
Ability to Disable Algorithms	To what degree are there controls in place to disable the algorithm if there is an issue with it?	1	2	3	4	5	
Panel B: ESG Instrument							
Factor Name	Key Evidence	Scoring					
CSR/ESG—Executive Focus	To what degree is there an executive strategic focus on CSR/ESG?	1	2	3	4	5	
CSR/ESG—Culture	To what degree is there a culture of ESG present in the Bank?	1	2	3	4	5	
CSR/ESG—Training	To what degree is there formal training required for ESG at the Bank?	1	2	3	4	5	
CSR/ESG—Environment	To what degree is there an organizational focus on business impact to the environment?	1	2	3	4	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	
CSR/ESG—Governance	To what degree is there an organizational focus on governance / compliance within the Bank?	1	2	3	4	5	

frameworks [4]. The following sections describe the instrument categories and provide rationale supported by references to the relevant literature for the components.

4.1 Organizational commitment

There are evolving models of how organizations will manage the collaboration of AI and human capabilities [36]. Understanding how to leverage RAI in an organization requires a broader integration of the social environment within which the AI operates [31]. Banks that possess high organizational commitment to RAI will have a Center of Excellence (COE) team focused on Responsible AI, allocate and measure significant financial investments, conduct formal training programs, incorporate a decision framework for AI, have meaningful engagement from top executives, and a focus on the culture of RAI [94]. There were elements of organizational commitment present in other summaries of RAI principles [43, 54, 62, 95], such as 'accountability', 'governance', 'culture', 'diversity', and 'humanity', which we incorporated in this new RAI principle and category to measure the focus of the organization [21, 106].

4.2 Transparency

Transparency (including interpretability and explainability) remains one of the most important areas within RAI governance [23, 31]. Firms are challenged to manage AI models that are explainable, interpretable, and understandable [18]. Banks that possess strong explainability and transparency will have a formal explainability governance process, a regulatory sandbox, model audit controls, model drift monitoring, potentially use of knowledge graphs and model cards, and lastly explore advanced ML techniques such as SHAP and LIME explainability. Transparency was present in each of the other referenced principles' summaries [43, 54, 62, 95].

Table 2 RAI instrument attributes

Organizational Commitment Attributes		
Factor Name	Rationale	Source
Organizational Commitment to RAI	**This category measures the degree of organizational commitment to RAI for accountability, governance, finance, and culture**	
Organizational Structure RAI Focus	Organizations that are committed to RAI will 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 [38]
Investment in RAI	To have such resources deployed to RAI, companies must commit significant investment	Borg [17]
ROI Analysis on RAI	Companies that are have high organizational commitment will align investments (e.g., corporate social responsibility and organizational purpose) with a positive financial return	Fraisse and Laporte [44]
Training for RAI	Corporations that are dedicated to RAI will also make investments in training for their employees to ensure that the general concepts are understood and AI is part of the working language of their business practices	Cihon et al. [31]
Culture of AI	Organizations that establish patterns and methods to nurture the culture to embrace RAI will have greater adoption and more effectiveness	Murphy and Largacha-Martínez [94]
Pareto Efficiency Frontier Decisions	Banks that embrace RAI 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. [83]
C-Suite Involvement	Companies that are serious about gaining competitive advantage from investments in RAI will have significant involvement from executive leadership with regular briefings and formal scorecards in place	Burkhardt et al. [21]
Transparency and Explainability Attributes		
Factor Name	Rationale	Source
Transparency & Explainability	**This category measures the degree of transparency in the AI in terms of the algorithms, and models that comprise the AI & ML**	
Explainability Governance	Model auditing through checklists, questionnaires, documentation, model card reporting have been employed as part of explainability and governance efforts	Ayling and Chapman [8]
Regulatory Sandbox (Visibility)	The term and capability of a 'regulatory sandbox' is gaining traction as a place where models can be demonstrated and approved	Goo and Heo [51]
Model Audit Controls	Model auditing is needed with advanced AI models (using unsupervised and reinforcement learning) because of their black box nature in which human operators do not fully understand how the AI derived the outcome	Adler et al. [3]
Model Drift Prevention Monitoring	Model drift is also important to explainability where various environmental factors may render some part of the ML lifecycle obsolete	Barros and Santos [10]
Use of Knowledge Graphs	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	Tiddi and Schlobach [121]
Use of Model Card Reporting	Model auditing through checklists, questionnaires, documentation, model card reporting have been employed as part of explainability and governance efforts	Mitchell et al. [88]
Advanced ML Explain (LIME, SHAP)	In the case of more advanced ML, different explanatory techniques are required with new tools such as LIME and SHAP	Gramegna and Giudici [52]

Table 2 (continued)

Fairness Attributes		
Factor Name	Rationale	Source
Fairness/Bias Mitigation	**This category measures the ability to mitigate bias in lending and prevent discrimination harm**	
Policy for Fairness in Models	Strong fairness policies consist of a review of guidelines, training, and executive reinforcement to engrain responsible and fair AI into company policies, guidelines, processes, and models	Lee and Floridi [72]
Fairness in Training Data	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	Hall et al. [55]
Human in the Loop	Research suggests it is important to have 'Human in The Loop' processes, which allows for a human to infuse context and judgement into decisions	Buckley et al. [19]
Legal Implications	Strong RAI/ML capabilities must focus on legal aspects, as the Bank is subject to regulation regarding the disparate impact of the policy	Cath [28]
Proxy Discrimination	There are also cases where the bias and discrimination are not specifically intended, and rather found incidentally through bias in training data or other data elements, which can lead to proxy discrimination	Prince and Schwarcz [99]
Model Reparation	Even with robust fairness policies and explainability capabilities in place models may still contain bias, thus action plans to detect and remediate the bias through model reparation must be present	Davis et al. [37]
Data Management Attributes		
Factor Name	Rationale	Source
Data Management & Quality	**This category measures the maturity of the data mgmt. processes that feeds the ML models**	
Data Privacy	Data privacy is one of the most important aspects of responsible data management	Stoyanovich et al. [119]
Differential Privacy Capability	Managing privacy while performing ML techniques that do not expose PII (personally identifiable information) is present in differential privacy capabilities	Dwork et al. [42]
EDA for Pre-modeling	A tool that is leveraged for advanced data management is EDA (exploratory data analysis), which provided a visual capability for analyzing the data	Wang [129]
Use of Data Pipeline Tools	Infusing responsible practices into the data pipeline toolset is critical for maintaining governance and controls for the data	Deepa and Ramesh [39]
Use of Big Data Lake	Many AI/ML environments will contain data lakes that store the data that can be used by the algorithms for training and execution	Martin [80]
CDO Involvement	Leadership support is key not only in the organizational commitment category, but critical to the success of responsible data management with CDO involvement	Helmy et al. [56]
Use of Synthetic Data	The notion of synthetic data also exists where the corporation will infuse manufactured data into the environment to influence the models	Campbell [26]
Use of DataOps	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. [110]
Right to be Forgotten	In addition to data privacy, Banks also enable the 'right to be forgotten', where a customer would like their data purged	Tjong Tjin Tai [122]

Table 2 (continued)

Security Attributes		
Factor Name	Rationale	Source
Security	**This category measures the specific data privacy and data security provisions are in place**	
Adversarial Cyber Attack Defense	A significant 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 et al. [6]
Data Encryption	One of the methods for ensuring data privacy and security is to apply encryption	Lauter [70]
Special Security Access for Production	Another method for ensuring data privacy and security access control to the databases and data	Wee and Nayak [130]
Security Processes for Unintended Usage	It is critical to have monitoring and controls in place to defend and remediate the system should an adversarial attack gains access to the ML models in production	Papernot [97]
Ability to Disable Algorithms	If the ML data pipeline or production environment has been compromised, it is important to have process and procedures in place to be able to disable the algorithms to limit impact in the environment	Verma et al. [126]

4.3 Fairness

Bias is inherent in human cognition and an unavoidable characteristic of data collected from human processes [41]. Bias manifests itself in scenarios where the result or action of a decision is perceived as unfair [7]. Though each of the following types of data are not equally relevant to credit underwriting; criminal records, bill payment history, education history, and residential address are examples of data that may contain or lead to bias [33]. As a result of this recognized potential for bias and discrimination [99], the regulatory agencies are focused on the Fair Credit Reporting Act (FCRA), and the Equal Credit Opportunity Act (ECOA). SR (Supervision and Regulation) 11–7 (Supervisory Guidance on Model Risk Management) from the Fed and Office of Comptroller of the Currency (OCC) explicitly defines the required risk management around credit lending. Banks concerned with fairness in credit lending will possess capabilities that focus on robust policy review, mitigating proxy discrimination, assessing training data management, understanding to what extent humans [58] are involved in the decision-making process, legal and regulatory considerations, and action plans for when the models and algorithms run amok. Fairness was also present in each of the other referenced summaries [43, 54, 62, 95].

4.4 Data management

AI systems use data that is generated through life, mirroring varied attributes which make it susceptible to containing bias [41]. Data privacy is one of the most important aspects of responsible data management [119]. Managing privacy while performing ML techniques that do not expose personally identifiable information (PII) is present in differential privacy capabilities [42]. Banks concerned with good data management ethics and governance will focus on data privacy as well as differential privacy, leverage data pipeline tools as well as data Ops, EDA for pre-modeling, knowledge graphs for visual representation of data, big data lakes, generating synthetic data (though nascent and sometimes debated) [26], enabling the right to be forgotten, and have Chief Data Officer (CDO) involvement [80]. In terms of other references to this principle and category, there was partial alignment with Jobin et al. [62] and Fjeld et al. [43] in terms of data privacy, partial alignment with Myers and Nejkov [95] who focused on data validation, and no clear alignment with Hagendorff [54].

4.5 Security

Elements of security to prevent intrusion and guard vulnerabilities are paramount to ensuring fairness, safety, and privacy in RAI [97]. In making the capabilities of AI explainable and transparent, maintaining user trust and individual privacy is critical [119]. Malicious actors may insert data into training or production environments, which could result in a data poisoning attack [60]. In terms of security elements, banks focus on capabilities that mitigate intrusions, ensure data encryption, enable processes for handling models should they become infected or fall into unintended possession, enable the ability to control the algorithms, and

Table 3 Proxy RAI instrument

(What is the study looking for?) Evidence of mature responsible AI	(How to score?) Coding mechanism	(What to look for?) Keywords/phrases same or similar in nature = 1	(What to look for?) Keywords/phrases same or similar in nature = .5	(What to look for?) keywords/phrases absence of keywords = 0
RAI principles Does the Bank have RAI Published Principles?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No Evidence
RAI COE Does the Bank have a RAI Center of Excellence	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No Evidence
Articles in the press Does the Bank have RAI Articles in the Press?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No Evidence
Investor relations focus Is RAI mentioned in 10 K / Shareholder letter?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No Evidence
Website presence Is there a prominent RAI link on Website?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No Evidence
Research partnerships Does the Bank have Research Partnerships with Universities?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No Evidence
Research Dept. Does the Bank have an Internal RAI Research Dept / Publications?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning Data Science	No evidence
RAI careers Does the Bank have Careers in RAI related field?	If RAI is present = 1; if AI is present = .5; else 0	Responsible AI or Ethical AI Explainability & Transparency AI Regulation or Credit Risk AI Responsible Business	Artificial Intelligence Machine Learning AI/ML coding languages, e.g. Python	No Evidence

Table 4 MTMM matrix

Instrument	RAI instrument	MTMM type	RAI proxy	MTMM type	ESG instrument	MTMM type
RAI proxy	0.882	Mono-trait—multi-method				
ESG instrument	0.553	Multi-trait—mono-method	0.398	Multi-trait—multi-method		
ESG sustainalytics	0.135	Multi-trait—multi-method	0.109	Multi-trait—multi-method	0.532	Mono-trait—multi-method

All correlations are based on the full sample ($n = 48$)

assign special security for production runtime environments. Security was present in the other referenced summaries [43, 54, 95], however, the concept was listed in terms of privacy in Jobin et al. [62].

5 Validity and reliability assessment

To advance the RAI instrument, the RAI trait and ESG trait are required to be represented from two different sets of data from different methods. Since there were no other extant RAI scores, we derived a secondary (proxy) instrument by coding publicly available archival data (Table 3). In addition, for purposes of testing discriminant validity in multiple ways we added an ESG trait survey panel to our instrument interviews to collect the ESG (instrument) score and additional ESG scores were obtained from Sustainalytics. Panel A collected data from the banks for the key categories and attributes of the RAI (instrument) score, and Panel B collected data for the ESG (instrument) score. As described, ensuring that the multi-traits of RAI and ESG were evaluated using the same method satisfied the multi-trait mono-method requirement.

5.1 Data description

The data sample for the instrument was collected from and represented by 48 of the 56 top US Banks listed in "*ADVRatings*—Top 50 Banks in America", representing more than 85% of the population of large banks and a significant portion of the credit lending (credit cards, mortgages, and auto loans). The top banks are typically public companies, highly regulated by the Fed and OCC and have ample resources to employ mature RAI. Banks are motivated on this topic as they field social pressure for fairness and represent a good current state proxy analysis on the industry's RAI capabilities. The data about the banks (collected in early 2022) was comprised of two RAI data elements (RAI (instrument) score and RAI (proxy) score) as well as two ESG data elements (ESG (Sustainalytics) score and ESG (instrument) score).

5.2 Validity

Measurement instruments can be evaluated for different kinds of validity. While face validity (which measures the degree of assessment effectiveness) and content validity (which measures the effectiveness of the construct) are supported by subjective evidence and argument, construct validity (which measures the effectiveness of the concept design) must be tested and confirmed empirically. We conducted validation interviews with over 40 banks' executives related to the MRM (Model Risk Manager) function to obtain rich insights to enhance the instrument. Interview questions, categories, and/or detailed measurement attributes in the instrument that were deemed irrelevant or had critical components missing were addressed in the final version of the instrument. Our instrument attained face validity and content validity through this process.

To test construct validity, we leveraged the multi-trait multi-method matrix approach introduced by Campbell and Fiske [25]. In this approach, the presence of construct validity (i.e., both convergent and discriminant validity) is observed if the following two conditions are satisfied:

(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 different constructs (i.e., multi-trait) computed through the same instrument (i.e., mono-method) exceeds the correlation between different constructs (i.e., multi-trait) scored through alternative instruments (i.e., multi-method).

To test for convergence and discriminant validity, we correlated the two RAI score (RAI (instrument) and RAI (proxy)) traits with two ESG score (ESG (instrument) and ESG (Sustainalytics)) traits. To satisfy the requirement for convergent validity, the RAI instrument must be significantly correlated with a conceptually similar construct (i.e., the RAI (proxy) score for the purposes of this study).

In the case of discriminant validity, the variables must *not* be as highly correlated with a seemingly related but conceptually different construct (e.g., ESG score of any format). The conceptually different construct can be measured with the same method (i.e., ESG (instrument) score) or a different method (i.e., ESG (Sustainalytics) score). In the first case, the test is for a similar method, but different traits, accomplished by correlating an RAI (instrument) score with an ESG (instrument) score. In the second case, the test is for different traits and different methods, accomplished by correlating an RAI (instrument) score with an ESG (Sustainalytics) score or correlating an RAI (proxy) score with an ESG (instrument) score. In addition to satisfy the second condition, the test is to confirm that the different traits using the same method correlate higher than different traits in different methods.

In the MTMM matrix in Table 4, the study displays correlations of RAI (instrument) scores and RAI (Proxy) scores (mono-trait multi-method), RAI (instrument) scores and ESG (Sustainalytics) scores (multi-trait multi-method), RAI (instrument) scores and ESG (instrument) scores (multi-trait mono-method), and ESG (instrument) scores and ESG (Sustainalytics) scores (mono-trait multi-method).

We find that the mono-trait multi-method correlation of the RAI (instrument) score and RAI (proxy) score is significantly high at ($r = 0.882$), demonstrating strong evidence of convergent validity. We then test the first condition and compare the mono-trait multi-method of (RAI (instrument) score and RAI (proxy) score) of ($r = 0.882$) with the multi-trait mono-method correlation of (RAI (instrument) score and ESG (instrument) score) of ($r = 0.553$) and satisfy a first case (i)(a) of discriminant validity. Next, we compare the mono-trait multi-method of (RAI (instrument) score and RAI (proxy) score) of ($r = 0.882$) with the multi-trait multi-method correlations of (RAI (Instrument) score and ESG (Sustainalytics) score) of ($r = 0.135$) and (RAI (proxy) score and ESG (instrument) score) of ($r = 0.398$) and satisfy a second case (i)(b) of discriminant validity. Lastly, we test the second condition (ii) and compare a case of multi-trait mono-method (RAI (instrument) score and ESG (instrument) score) of ($r = 0.553$) with multi-trait multi-method (RAI (instrument) score and ESG ((Sustainalytics) score) of ($r = 0.135$) and (RAI (proxy) score and ESG (instrument) score) of ($r = 0.398$), comprehensively satisfying the criteria for discriminant validity. The MTMM analysis demonstrates clear evidence of construct validity (concept design accuracy), highlighting both convergent validity (which measures how closely a test is related to other tests of the same construct) as well as discriminant validity (which measures the extent to which a test is not related to other tests of a different construct) of the RAI instrument.

5.3 Reliability

With the RAI instrument receiving support for convergent and construct validity, we utilized two different statistical techniques to ensure that the RAI instrument contained the most relevant measurement elements and satisfied internal consistency reliability.

First, we computed Cronbach's alpha (which measures the internal consistency of items in the survey scale) on the RAI instrument with a goal value of $\alpha > 0.7$. This test was conducted on each of the five instrument categories to assess the degree of cohesion of the attributes of each category. For the RAI instrument data, Cronbach's alpha for Organizational Commitment (containing seven items) is 0.898, Explainability (containing seven items) is 0.951, Fairness (containing six items) is 0.931, Data Management (containing nine items) is 0.947, and Security (containing five items) is 0.892 providing significant support of internal consistency reliability.

Second, we employed confirmatory factor analysis (CFA) (which measures how well the variables represent the number of constructs) with a goal factor loading threshold of 0.5 [75]. We computed the CFA for the entire RAI instrument data as well. The CFA for the 31 factors ranges from 0.638 to 0.873 with all factors exceeding the 0.5 loading threshold. Overall, the scores from both Cronbach's alpha as well as CFA statistics offer convincing evidence of internal consistency reliability of our RAI measurement instrument.

For the RAI proxy measure, we engaged two raters and provided instructions to each rater separately to interpret the data and record 1, 0.5, or 0 if evidence is found for the attribute. Each rater coded 384 items (i.e., 8 items per bank for 48 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, 0.5, or 0 accordingly. The coding value of 1 was recorded when the full criteria of RAI was met; the coding value of 0.5 was recorded when AI was present, but RAI was not; the coding value of 0 was recorded when neither RAI nor AI was present. The initial inter-rater agreement was 97.7%, and even after accounting for chance correlation between our two different raters [34], the Cohen's kappa coefficient (which measures the inter-rater reliability for categorical items) is 0.965 ($p < 0.001$). A Cohen's kappa coefficient above 0.60 shows acceptable levels of inter-rater reliability [69]. The

very high kappa coefficient achieved in this study provides strong evidence for inter-rater reliability of the RAI proxy measure.

6 Discussion and conclusions

AI is a powerful and rapidly evolving technology that many corporations are adopting, creating a race between reaping the benefit of the capability and addressing the RAI governance for fair AI deployment [45]. RAI has also become a critical topic over the past decade in conjunction with a focus on responsible business [68], and a focus on DE&I (diversity, equity, and inclusion) as well as fairness driven by bank's ESG (environmental, social, governance) agendas [85, 86]. Due to the profitability of credit lending (mortgages, auto loans, and credit cards) for banks, there is a constant push for innovation and efficiency [1, 13]. This study has addressed a gap in the industry by inventing a new measurement instrument (RAI) with which banks can now assess the maturity of their RAI capabilities.

6.1 Theoretical contributions

The key contribution of this paper is the introduction of a statistically valid and reliable RAI measurement instrument that organizations can deploy to assess the RAI maturity in their AI capabilities. The study incorporated the following categories (explainability, fairness, data management, and security) based on the 'referenced ranking score' from the analysis of RAI principles, which was a comprehensive review of the major RAI references from Jobin et al. [62], Hagendorff [54], Myers and Nejkov [95] and Fjeld et al. [43] as well as published RAI principles from 33 AI focused organizations listed in the Appendix.

Notably, as a second contribution, we add a novel RAI principle and category named 'organizational commitment', which incorporates elements of accountability, culture, strategy, investment, and decision-making, which we believe is paramount for organizations in their quest to leverage ethical algorithms to develop mature RAI capabilities [82, 90, 106]. This new RAI principle as an instrument category contributes to the literature, as it brings focus to leadership around ethics in technology and a commitment to RAI, amplifying extant principles of accountability and governance [21].

A third contribution we make is the creation of an additional assessment tool (RAI (proxy) instrument) which reviews key indicators of leadership focus on RAI. This contributes to the literature, as this additional instrument enables a different method of assessing RAI maturity. The RAI (proxy) instrument also lends itself to the ability to scale widely, since it is based on publicly available archival data, automated context analysis tools could scour the internet and assess various companies for their RAI maturity.

Lastly, a fourth contribution is the distinction between RAI and ESG as reported by the results of the MTMM analysis. RAI and ESG are seemingly related and in fact researchers may argue that RAI is part of ESG [85, 86]. RAI and ESG do have a shared component in that they are both increasingly subject to formal regulation imparting legal implications for firms that deviate from the requirement [22, 28], however, there is a clear distinction in our research. Through the validity tests of the instrument and the initiative of measuring ESG in the same method demonstrating relatively lower correlation between the (RAI (instrument) score and the ESG (instrument) score) (leveraging multi-trait mono-method) compared with the (RAI (instrument) score and the RAI (proxy) score), we establish a clear contrast between RAI maturity and ESG. In our research, we found that nearly all banks had a published ESG report, however, only a handful had their RAI principles published. The contribution of this potential paradox highlighted in our findings of this special intersection may motivate researchers to explore the relatedness of RAI and ESG in more detail.

6.2 Applied implications

With the introduction of the RAI instrument, banks can assess the degree of maturity they possess in their RAI. There are multiple implications that are enabled by the ability to measure RAI, especially considering recent innovation around generative AI exemplifying the power that the AI technology possesses. First, advancement of this new instrument will enable banks to highlight investments and capabilities in their RAI programs for customer acquisition. This implication is important as in the context of financial borrowers seeking an unbiased and fair credit lending process, the bank may use the RAI maturity score to craft positive messaging about fairness in lending in their marketing and advertising.

Second, we posit that RAI will increase in importance due to additional regulatory governance over AI usage and considerations for liability [22, 27, 124, 128]. There is broad speculation that formal regulation will be introduced to mandate transparency in the form of explainability and model auditing processes (i.e. Algorithm Audit) [66]. We do not assert that banks have ignored the ethical risks to date and in fact the banks have attempted to create bias free application processes and credit worthiness evaluation data [72]. However, the introduction of alternate data increases potential for bias due to the additional data attributes that are considered [61]. This raises a question on whether RAI is merely forcing banks further regulatory consideration. While the traditional scoring has been efficient for the credit process, we argue that

RAI is indeed advocating additional transparency through explainability, thus illuminating the discussion [27, 78]. In addition, with momentum building for the 2022 Algorithmic Accountability Act sponsored by US Senators (Wyden and Booker), and other regulatory measures such as "truth, fairness and equity in AI" [22], banks would be well advised to proactively build these explanatory capabilities for regulatory requests. The RAI instrument can serve as a communication tool for banks and regulators to align on maturity assessments and action plans to enhance fairness in credit lending.

Third, there are a few potential areas of implications for bank stakeholders that could benefit from leveraging the RAI instrument. Similar to ESG and CSR [50], RAI could influence how both institutions and individuals invest [85, 86]. RAI could also impact how investment research analysts write about the stocks because environmental and social impacts are expected to influence stakeholders. In terms of fairness and SRI (socially responsible investing) principles, there seems to be synergies with the conceptual nature of how social issues can impact investments [46]. Following ESG precedence, investors could research the RAI score for a bank to determine worthiness. In addition, the analysts who cover stocks could leverage the RAI to improve accountability. From a strategic growth perspective, focusing on fairness in credit lending enables the bank to leverage the RAI assessment score to advertise alignment with ESG and CSR statements.

Fourth, the implication for the borrowers is significant, as with the industrialization of an RAI instrument, the banks can both comply with regulatory definitions for fairness in lending [114], as well as optimize profitability for the bank, resulting in more borrowers receiving loans. Making decisions responsibly is key to future leadership [102], and the ability to assess the RAI capabilities will serve executives well.

Lastly, we employed a CMM (capability maturity model) analysis to generalize the maturity of the banking industry and reported a mean of 53.74% across the distribution of banks. The CMM model had the most frequency in the "Operational" level of maturity, which was described by Gartner and Panetta [48] as "AI in production, creating value by e.g., processing optimization or product/service innovations". With AI becoming ever more pervasive, leadership decisions for responsible business will increasingly be data driven through AI [118]. We argue that there will be significant investment by banks to improve business process efficiency and productivity through the AI, in turn driving a continued focus on RAI [49]. The evidence provided through this lens of the data illustrates the state of maturity of RAI in the Banking industry in 2022 creating a call to action for banks to focus on enhancing RAI maturity.

6.3 Limitations and future research

We develop a novel RAI instrument as well as leverage the instrument to assess the maturity score of RAI capabilities in banks. The survey was customized for the Banking industry since the validity and reliability tests were done using banking industry survey and archival data. Since we review the RAI programs specifically regarding banks' credit lending, we particularly focus on the ML of the credit underwriting algorithms, models, and data that are associated with credit lending decisions. This is a limitation since in its current form, the instrument is not generalizable to other industries, however, with some minor modifications, the instrument could be made more generic or could be tailored to other industries, as some instruments of different nature originate in this capacity.

While we conducted a comprehensive review of the RAI principles included in the instrument including a careful review of Jobin et al. [62], Hagendorff [54], Myers and Nejkov [95] and Fjeld et al. [43], it is certainly possible that other principles could be incorporated that are deemed more relevant to a tool of this nature.

The RAI instrument survey collected data from the MRM (Model Risk Manager) bank executives as a self-assessment score and not a researcher interrogation of the actual environment, thus this could be perceived as a limitation. In addition, a similar limitation exists in the Proxy RAI instrument, as the study collected data from public archival data into the eight categories the corresponding researcher chose. It is possible that there could have been different categories to record the data.

Another potential area for critique of the instrument is the scoring calculation we designed. Researchers may debate that some principles are more important than others. Transparency and explainability may be deemed the most important instrument category and deserve to be weighted higher [20, 35], and others may contend that the 'organizational commitment' category should bear more weight [21, 106]. Our testing did not find evidence for weighting one category or attribute more heavily than another, therefore, we concluded that the RAI instrument should maintain equal weighting for all the categories and attributes.

The future research opportunities for this RAI instrument are significant. First, this RAI instrument could be generalized outside of banking. Second, this RAI instrument score could become a standard independent variable for future research to predict other aspects of companies, for example a correlation with corporate financial metrics, ESG-CSR scores [104], brand reputation indices [101], or TMT diversity [107]. Third, there could be studies on how AI is impacting decision-making in terms of capability

investment. For example, reviewing how banks invest in fraud detection AI capabilities as compared with how they invest in RAI capabilities.

Finally, if responsibility becomes a significant measurement for investing, partnering, or buying from responsible firms, RAI 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 is underway with next generation firms [87, 118].

Appendix 1

Representation of RAI principles research distribution

Name	RAI principles in focus
Accenture	(1) Fairness (2) Accountability (3) Transparency (4) Explainability (5) Privacy
AI4People	(1) Beneficence: promoting well-being, preserving dignity, and sustaining the planet (2) Non-maleficence: privacy, security and "capability caution" (3) Autonomy: the power to decide (whether to decide) (4) Justice: promoting prosperity and preserving solidarity (5) Explicability: enabling the other principles through intelligibility and accountability
ALTAI (EU—European Commission's High-Level Expert Group)	(1) Human agency and oversight (2) Technical robustness and safety (3) Privacy and data governance (4) Transparency (5) Diversity, non-discrimination and fairness (6) Societal and environmental well-being (7) Accountability
Amazon	(1) Fairness (2) Accountability (3) Transparency (4) Ethics
Asilomar principles	(1) Safety (2) Failure Transparency (3) Judicial Transparency (4) Responsibility (5) Value (Human) Alignment (6) Human Values (7) Personal Privacy (8) Liberty and Privacy (9) Shared Benefit (10) Shared Prosperity (11) Human Control (12) Non-subversion (13) AI (mitigate) Arms Race
Bain	(1) Well-being & Dignity (2) Privacy (3) Fair treatment (4) Data transparency, accuracy (5) Reliability, Robustness (Testing of Models)
BCG	(1) Accountability (2) Transparency and "explainability" (3) Fairness and equity (4) Safety, security, and robustness (5) Data and privacy governance (6) Social and environmental impact mitigation (7) Human plus AI
Cap Gemini	(1) Carefully delimited impact (2) Sustainable (3) Fair (4) Transparent and explainable (5) Controllable with clear accountability (6) Robust and safe (7) Respectful of privacy and data protection
Deloitte	(1) Fair and impartial (2) Transparent and explainable (3) Responsible and accountable (4) Robust and reliable (5) Respectful of privacy (6) Safe and secure
Department of defense	(1) Alignment of expectations (2) Fairness (3) Accountability (4) Transparency (5) Mitigate harm
EY	(1) Fairness (2) Reliability (3) Explainability (4) Ethics
Facebook	(1) Privacy & Security (2) Fairness & Inclusion (3) Robustness & Safety (4) Transparency & Control (5) Accountability & Governance

Name	RAI principles in focus
Forbes	(1) Accountable (2) Impartial (3) Resilient (4) Transparent (5) Secure (6) Governed
Google	(1) Be socially beneficial (2) Avoid creating or reinforcing unfair bias (3) Be built and tested for safety (4) Be accountable to people (5) Incorporate privacy design principles (6) Uphold high standards of scientific excellence (7) Be made available for uses that accord with these principles
(IEAIML) Institute for Ethical AI and Machine Learning	(1) Human augmentation (2) Bias evaluation (3) Explainability by justification (4) Reproducible operations (5) Displacement strategy (6) Practical accuracy (7) Trust by privacy (8) Security risks
IBM	(1) Explainability (2) Fairness (3) Robustness (4) Transparency (5) Privacy
IEEE	(1) Human Rights (2) Well-being (3) Accountability (4) Transparency (5) Extending benefits and minimizing risks of misuse
IICP	(1) Collaboration (2) Transparency (3) Controllability (4) Safety (5) Security (6) Privacy (7) Ethics (8) User assistance (9) Accountability
Informatica	(1) Fair and equitable (2) Social ethics (3) Accountability and responsibility (4) Systemic transparency (5) Data and AI governance (6) Interpretability and explainability
KPMG	(1) Prepare employees now (2) Develop strong oversight and governance (3) Align cybersecurity and ethical AI (4) Mitigate bias (5) Increase transparency
McKinsey	(1) Appropriate data acquisition (2) Data-set suitability (3) Fairness of AI outputs (4) Regulatory compliance and engagement (5) Explainability
Microsoft	(1) Fairness (2) Reliability & Safety (3) Privacy & Security (4) Inclusiveness (5) Transparency (6) Accountability
Medium	(1) Fairness (2) Transparency (3) Empathy (4) Robustness
OECD	(1) Inclusive growth, sustainable development and well-being (2) Human-centered values and fairness (3)Transparency and explainability (4) Robustness, security and safety (5) Accountability
Partnership for AI	(1) Fairness (2) Transparency (3) Accountability (4) Human rights (5) Safety
PWC	(1) Interpretability (2) Reliability and robustness (3) Security (4) Accountability (5) Beneficiality (6) Privacy (7) Human agency (8) Lawfulness (9) Fairness (10) Safety
Responsible AI Institute	(1) Accountability (2) Bias & Fairness (3) Consumer Protection (4) Robustness, Security, Safety (5) Explainability & Interpretability (6) Systems Operations
Salesforce	(1) Being of benefit (2) Human value alignment (3) Open debate between AI researchers and policymakers (4) Cooperation, trust and transparency in systems (5) Safety and Responsibility
Stanford	(1) Human-centered (2) Fairness/bias mitigation (3) Do Good/Do no Harm (4) Support diversity in AI

Name	RAI principles in focus
Telefonica	(1) Fair AI (2) Transparent & Explainable AI (3) Human-centric AI (4) Privacy & Security by Design (5) Third Parties
Turing Institute (Leslie)	(1) Fairness (2) Accountability (3) Sustainability (4) Safety (5) Transparency
Twitter	(1) Taking responsibility for our algorithmic decisions (2) Equity and fairness of outcomes (3) Transparency about our decisions and how we arrived at them (4) Enabling agency and algorithmic choice
World Economic Forum	(1) Active inclusion (2) Fairness (3) Right to Understanding (4) Access to redress

Acknowledgements We would like to thank Theresa Lant and Mazhar Islam as well as the anonymous reviewers and the editor for their thoughtful comments on earlier drafts of this manuscript.

Data availability Data and statistics used for the study will be made available upon request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. This research was conducted on the corresponding author's own time as part of his Pace University doctoral dissertation and was not affiliated in any capacity with his employer, Accenture.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Abedifar, P., Molyneux, P., Tarazi, A.: Non-interest income and bank lending. J. Bank. Finance **87**, 411–426 (2018)
2. Adam, M., Wessel, M., Benlian, A.: AI-based chatbots in customer service and their effects on user compliance. Electron. Mark. **31**, 427–445 (2021)
3. Adler, P., Falk, C., Friedler, S.A., Nix, T., Rybeck, G., Scheidegger, C., Venkatasubramanian, S.: Auditing black-box models for indirect influence. Knowl. Inform. Syst. **54**(1), 95–122 (2017)
4. AIEthicist.: AI frameworks, guidelines, toolkits. AI Frameworks. Retrieved from https://www.aiethicist.org/frameworks-guidelines-toolkits (2021)
5. Ameen, N., Tarhini, A., Reppel, A., Anand, A.: Customer experiences in the age of artificial intelligence. Comput. Hum. Behav. **114**, 1–12 (2021)
6. Anthi, E., Williams, L., Rhode, M., Burnap, P., Wedgbury, A.: Adversarial attacks on machine learning cybersecurity defences in industrial control systems. J. Inform. Secur. Appl. **58**, 1–8 (2021)
7. Arnold, T., Scheutz, M.: The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. Ethics Inf. Technol. **20**(1), 59–69 (2018)
8. Ayling, J., Chapman, A.: Putting AI ethics to work: are the tools fit for purpose? AI Ethics **2**, 405–429 (2022)
9. Babic, B., Chen, D.L., Evgeniou, T., Fayard, A.L.: A better way to onboard AI. Harv. Bus. Rev. **98**(4), 56–65 (2021)
10. Barros, R.SMd., Santos, S.GTd.C.: An overview and comprehensive comparison of ensembles for concept drift. Inform. Fusion **52**, 213–244 (2019)
11. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Zhang, Y.: AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating algorithmic bias. IBM J Res Develop. 63(4/5), 4:1–4:15 (2019)
12. Bertsimas, D., Farias, V.F., Trichakis, N.: On the efficiency-fairness trade-off. Manage. Sci. **58**(12), 2234–2250 (2012)
13. Biswas, S., Carson, B., Chung, V., Singh, S., Thomas, R.: AI-bank of the future: can banks meet the AI challenge? McKinsey & Company **1**(2020), 1–14 (2020)
14. Boddington, P.: Toward a code of ethics for artificial intelligence. Springer, Cham (2017)
15. Boddington, P., Millican, P., Wooldridge, M.: Minds and machines special issue: ethics and artificial intelligence. Mind. Mach. **27**(4), 569–574 (2017)
16. Bolton, R.J., Hand, D.J.: Statistical fraud detection: a review. Inst. Math. Stat. **17**(3), 235–249 (2002)
17. Borg, J.S.: Four investment areas for ethical AI: transdisciplinary opportunities to close the publication-to-practice gap. Big Data Soc. **8**(2), 1–4 (2021)
18. Boza, P., Evgeniou, T.: Implementing Ai principles: frameworks, processes, and tools. INSEAD Working Paper No. 2021/04/DSC/TOM (2021)
19. Buckley, R.P., Zetzsche, D.A., Arner, D.W., Tang, B.W.: Regulating artificial intelligence in finance: putting the human in the loop. Syd. Law Rev. **43**(1), 43–81 (2021)
20. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. **70**, 245–317 (2021)
21. Burkhardt, R., Hohn, N., & Wigley, C.: Leading your organization to responsible AI. McKinsey Analytics, May, 1–8. Retrieved from: https://www.mckinsey.com/capabilities/quantumblack/our-insights/leading-your-organization-to-responsible-ai (2019)
22. Burt, A.: New AI regulations are coming. Is your organization ready? Harvard Business Review, April 30. Retrieved from: https://hbr.org/2021/04/new-ai-regulations-are-coming-is-your-organization-ready (2021)
23. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable machine learning in credit risk management. Comput. Econ. **57**(1), 203–216 (2020)
24. Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., Holzinger, A.: Quod erat

demonstrandum?—towards a typology of the concept of explanation for the design of explainable AI. Expert Syst. Appl. (2023). https://doi.org/10.1016/j.eswa.2022.118888
25. Campbell, D.T., Fiske, D.W.: Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull. **56**(2), 81–105 (1959)
26. Campbell, M.: Synthetic data: how AI is transitioning from data consumer to data produce and why thats important. Computer **52**(10), 89–91 (2019)
27. Candelon, F., Carlo, R.C.D., Bondt, M.D., Evgeniou, T.: AI regulation is coming. Harv. Bus. Rev. **99**(5), 102–113 (2021)
28. Cath, C.: Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos. Transact. R. Soc. A: Math. Phys. Eng. Sci. 376(2133) (2018)
29. Cavello, B.: 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/ (2020). Accessed 12 Apr 2022
30. Cheng, L., Varshney, K.R., Liu, H.: Socially responsible AI algorithms: Issues, purposes, and challenges. J. Artif. Intell. Res. **71**, 1137–1181 (2021)
31. Cihon, P., Schuett, J., Baum, S.D.: Corporate governance of artificial Intelligence in the public interest. Information **12**(7), 1–30 (2021)
32. Coates, D. L., Martin, A.: An instrument to evaluate the maturity of bias governance capability in artificial intelligence projects. IBM J. Res. Develop. 63(4/5), 7:1–7:15 (2019)
33. Coeckelbergh, M.: AI ethics. MIT Press, Cambridge (2020)
34. Cohen, J.: A coefficient of agreement of nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
35. Cortese, J.F.N.B., Cozman, F.G., Lucca-Silveira, M.P., et al.: Should explainability be a fifth ethical principle in AI ethics? AI Ethics (2022). https://doi.org/10.1007/s43681-022-00152-w
36. Daugherty, P.R., Wilson, H.J.: Radically human. Harvard Business Review Press, Boston (2022)
37. Davis, J.L., Williams, A., Yang, M.: Algorithmic reparation. Big Data Soc. **8**(2), 1–12 (2021)
38. de Laat, P.B.: Companies committed to responsible AI: from principles towards implementation and regulation? Philos. Technol. **34**(4), 1135–1193 (2021)
39. Deepa, B., Ramesh, K.: Production level data pipeline environment for machine learning models. Paper presented at the 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Tamil Nadu, India (2021)
40. Dhal, P., Azad, C.: A comprehensive survey on feature selection in the various fields of machine learning. Appl. Intell. **52**, 4543–4581 (2022)
41. Dignum, V.: Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer, Cham (2019)
42. Dwork, C., Rothblum, G. N., & Vadhan, S.: Boosting and differential privacy. Paper presented at the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science, Las Vegas, NV (2010)
43. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society (white papers). Retrieved from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420 (2020). Accessed 1 May 2022
44. Fraisse, H., Laporte, M.: Return on investment on artificial intelligence: the case of bank capital requirement. J. Bank. Finance **138**, 1–16 (2022)
45. Fu, R., Aseri, M., Singh, P.V., Srinivasan, K.: "Un"fair machine learning algorithms. Manage. Sci. **68**(6), 4173–4195 (2022)
46. Gadhoum, Y.: Artificial intelligence trends and ethics: issues and alternatives for investors. Intell. Control. Autom. **13**(1), 1–15 (2022)
47. Gallego-Gomez, C., De-Pablos-Heredero, C.: Artificial intelligence as an enabling tool for the development of dynamic capabilities in the banking industry. Int. J. Enterp. Inf. Syst. **16**(3), 20–33 (2020)
48. Gartner, & Panetta, K.: The CIO's guide to artificial intelligence. Retrieved from: https://www.gartner.com/smarterwithgartner/the-cios-guide-to-artificial-intelligence (2019). Accessed 12 Apr 2022
49. Ghosh, B., Prasad, R., Pallail, G.: The automation advantage: embrace the future of productivity and improve speed, quality, and customer experience through AI. McGraw Hill, New York (2021)
50. Gillan, S.L., Koch, A., Starks, L.T.: Firms and social responsibility: a review of ESG and CSR research in corporate finance. J. Corp. Finance (2021). https://doi.org/10.1016/j.jcorpfin.2021.101889
51. Goo, J.J., Heo, J.-Y.: The Impact of the regulatory sandbox on the fintech industry, with a discussion on the relation between regulatory sandboxes and open innovation. J. Open Innov. Technol Market Complexity **6**(2), 43–61 (2020)
52. Gramegna, A., Giudici, P.: SHAP and LIME: an evaluation of discriminative power in credit risk. Front. Artif. Intell. **4**, 140–146 (2021)
53. Haenlein, M., Kaplan, A.: A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. Calif. Manage. Rev. **61**(4), 5–14 (2019)
54. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020)
55. Hall, P., Cox, B., Dickerson, S., Ravi Kannan, A., Kulkarni, R., Schmidt, N.: A United States fair lending perspective on machine learning. Front. Artif. Intell. **4**, 1–9 (2021)
56. Helmy, M., Mazen, S., Helal, I.M., Youssef, W.: Analytical study on building a comprehensive big data management maturity framework. Int. J. Inform. Sci. Manag. **20**(1), 225–255 (2022)
57. Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H.: Improving fairness in machine learning systems. Proceedings of the CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, (paper#600) 1–16 (2019)
58. Holzinger, A., Plass, M., Kickmeier-Rust, M., et al.: Interactive machine learning: experimental evidence for the human in the algorithmic loop. Appl. Intell. **49**, 2401–2414 (2019)
59. Hunkenschroer, A.L., Luetge, C.: Ethics of AI-enabled recruiting and selection: a review and research agenda. J. Bus. Ethics **178**(4), 977–1007 (2022)
60. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B.: Manipulating machine learning: poisoning attacks and countermeasures for regression learning. Paper presented at the 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA (2018)
61. Jagtiani, J., Lemieux, C.: The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. Financ. Manage. **48**, 1009–1029 (2019)
62. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
63. Kavanagh, C.: New tech, new threats, and new governance challenges: an opportunity to craft smarter responses? Carnegie Endowment for International Peace. Retrieved from: https://carnegieendowment.org/2019/08/28/new-tech-new-threats-and-new-governance-challenges-opportunity-to-craft-smarter-responses-pub-79736 (2019). Accessed 1 May 2022

64. Kelley, S.: Employee perceptions of the effective adoption of AI principles. J. Bus. Ethics **178**(4), 871–893 (2022)
65. Kinkel, S., Baumgartner, M., Cherubini, E.: Prerequisites for the adoption of AI technologies in manufacturing—evidence from a worldwide sample of manufacturing companies. Technovation (2022). https://doi.org/10.1016/j.technovation.2021.102375
66. Koshiyama, A., Kazim, E., Treleaven, P.: Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI ML and associated algorithms. Computer **55**(4), 40–50 (2022)
67. Kumar, I. E., Hines, K. E., Dickerson, J. P.: Equalizing credit opportunity in algorithms: aligning algorithmic fairness research with U.S. fair lending regulation. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 357–368, New York: Association for Computing Machinery (2022)
68. Lacy, P., Long, J., Spindler, W.: The circular economy handbook: realizing the circular advantage. Palgrave Macmillan, London (2020)
69. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
70. Lauter, K.: Private AI: machine learning on encrypted data. In: Chacón Rebollo, T., Donat, R., Higueras, I. (eds.) Recent advances in industrial and applied mathematics, SEMA SIMAI Springer Series, 1. Springer, Cham (2022)
71. Langenbucher, K.: Responsible A.I. credit scoring: a legal framework. Eur. Law Rev. **25**, 527–572 (2020)
72. Lee, M., Floridi, L.: Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Mind. Mach. **31**, 165–191 (2020)
73. Leo, M., Sharma, S., Maddulety, K.: Machine learning in banking risk management: a literature review. Risks **7**(1), 1–22 (2019)
74. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 1–45 (2020)
75. Long, J.: Confirmatory factor analysis. Sage Publications, Thousand Oaks (1983)
76. Loureiro, S.M.C., Guerreiro, J., Tussyadiah, I.: Artificial intelligence in business: state of the art and future research agenda. J. Bus. Res. **129**, 911–926 (2021)
77. Lu, N., Zhang, G., Lu, J.: Concept drift detection via competence models. Artif. Intell. **209**, 11–28 (2014)
78. MacCarthy, M.: AI needs more regulation, not less. Retrieved from: https://www.brookings.edu/research/ai-needs-more-regulation-not-less/ (2020). Accessed 12 Apr 2022
79. Martín, A., Fernández-Isabel, A., Martín de Diego, I., Beltrán, M.: A survey for user behavior analysis based on machine learning techniques: current models and applications. Appl. Intell. **51**, 6029–6055 (2021)
80. Martin, K.: Ethical issues in the big data industry. MIS Q. Exec. **14**(2), 67–85 (2015)
81. Martin, K.: Ethical implications and accountability of algorithms. J. Bus. Ethics **160**(4), 835–850 (2018)
82. Martin, K.: Designing ethical algorithms. MIS Q. Executive **18**(2), 129–142 (2019)
83. Martinez, N., Bertran, M., & Sapiro, G.: Minimax Pareto fairness: a multi objective perspective. International Conference on Machine Learning: Proceedings of Machine Learning Research, 119, 6755-6764 (2020)
84. McCanless, M.: Banking on alternative credit scores: auditing the calculative infrastructure of U.S. consumer lending. Econ. Space, 1–19 (2023)
85. Minkkinen, M., Niukkanen, A., Mäntymäki, M.: What about investors? ESG analyses as tools for ethics-based AI auditing. AI Soc (2022). https://doi.org/10.1007/s00146-022-01415-0
86. Minkkinen, M., Zimmer, M.P., Mäntymäki, M.: Co-shaping an ecosystem for responsible AI: five types of expectation work in response to a technological frame. Inf. Syst. Front. (2022). https://doi.org/10.1007/s10796-022-10269-2
87. Miska, C., Mendenhall, M.E.: Responsible leadership: a mapping of extant research and future directions. J. Bus. Ethics **148**(1), 117–134 (2015)
88. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Gebru, T.: Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA (2019)
89. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019)
90. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2), 1–21 (2016)
91. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**(4), 2141–2168 (2020)
92. Moscato, V., Picariello, A., Sperlí, G.: A benchmark of machine learning approaches for credit score prediction. Expert Syst. Appl. (2021). https://doi.org/10.1016/j.eswa.2020.113986
93. Munoko, I., Brown-Liburd, H.L., Vasarhelyi, M.: The ethical implications of using artificial intelligence in auditing. J. Bus. Ethics **167**(2), 209–234 (2020)
94. Murphy, J.W., Largacha-Martínez, C.: Is it possible to create a responsible AI technology to be used and understood within workplaces and unblocked CEOs' mindsets? AI & Soc. (2021). https://doi.org/10.1007/s00146-021-01316-8
95. Myers, G., Nejkov, K.: Developing artificial intelligence sustainably: toward a practical code of conduct for disruptive technologies. EM Compass **80**, 1–8 (2020)
96. O'Neil, C.: Weapons of math destruction. Broadway Books, New York (2016)
97. Papernot, N.: A marauder's map of security and privacy in machine learning. Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security. New York: Association for Computing Machinery (2018)
98. Polyzotis, N., Roy, S., Whang, S.E., Zinkevich, M.: Data lifecycle challenges in production machine learning: a survey. SIGMOD Record **47**(2), 17–28 (2018)
99. Prince, A.E.R., Schwarcz, D.: Proxy discrimination in the age of artificial intelligence and big data. Iowa Law Rev. **105**(3), 1257–1318 (2020)
100. Rahman, N., Blake, L.: A review of CSR classification schemes and the operationalization of bolted-on vs. built-in CSR. Bus. Ethics Environ. Responsib. **30**(3), 248–261 (2021)
101. Rahman, N., Post, C.: Measurement issues in environmental corporate social responsibility (ECSR): toward a transparent, reliable, and construct valid instrument. J. Bus. Ethics **105**(3), 307–319 (2011)
102. Rahman, N., & De Feis, G.L.: Strategic decision-making: Models and methods in the face of complexity and time pressure. J. General Manag. **35**(2), 43–59 (2009)
103. Rahman, N., & Starbuck, W.H.: European and North American origins of competitive advantage. Adv. Strat. Manag. **27**, 313–351 (2010)
104. Rai, A.: Explainable AI: from black box to glass box. J. Acad. Mark. Sci. **48**(1), 137–141 (2019)
105. Rakova, B., Yang, J., Cramer, H., Chowdhury, R.: Where responsible AI meets reality: practitioner perspectives on enablers for shifting organizational practices. Assoc. Comput. Mach. **1**(1), 23 (2021)
106. Ransbotham, S., Khodabandeh,S., Fehling, R., LaFountain, B., & Kiron, D.: Winning with AI. MIT Sloan Management

Review and Boston Consulting Group. Retrieved from: https://sloanreview.mit.edu/projects/winning-with-ai/. (2019). Accessed 1 May 2022
107. Ratzan, J., Lant, T.: Top management team diversity in financial services: the influence of functional and demographic diversity on firm financial performance. Glob. J. Manag. Mark. **3**(1), 105–123 (2019)
108. Robertson, J., Diab, A., Marin, E., Nunes, E., Paliath, V., Shakarian, J., Shakarian, P.: Darknet mining and game theory for enhanced cyber threat intelligence. Cyber Defense Rev. **1**(2), 95–122 (2016)
109. Rodriguez, L.: All data is not credit data. Columbia Law Rev. **120**(7), 1843–1884 (2020)
110. Rodriguez, M., de Araújo, L. J. P., Mazzara, M.: Good practices for the adoption of dataops in the software industry. J. Phys. Conf. Ser. 1694 (2020)
111. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019)
112. Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D.E., Li, Y.: An explainable AI decision-support-system to automate loan underwriting. Expert Syst. Appl. (2020). https://doi.org/10.1016/j.eswa.2019.113100
113. Satell, G., & Abdel-Magied, Y.: AI fairness isn't just an ethical issue. Harv. Bus. Rev. October. Retrieved from: https://hbr.org/2020/10/ai-fairness-isnt-just-an-ethical-issue (2020)
114. Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., Liu, Y.: How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. Artif. Intell. https://doi.org/10.1016/j.artint.2020.103238 (2020)
115. Schiff, D., Rakova, B., Ayesh, A., Fanti, A., Lennon, M.: Explaining the principles to practices gap in AI. IEEE Technol. Soc. Mag. **40**(2), 81–94 (2021)
116. Shao, Z., Zhao, R., Yuan, S., Ding, M., Wang, Y.: Tracing the evolution of AI in the past decade and forecasting the emerging trends. Expert Syst. Appl. (2022). https://doi.org/10.1016/j.eswa.2022.118221
117. Singh, P.J., Franceschini, F., Smith, A.: An empirically validated quality management measurement instrument. Benchmarking **13**(4), 493–522 (2006)
118. Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., Machtynger, L.: Artificial intelligence (AI) in strategic marketing decision-making: a research agenda. Bottom Line **33**(2), 183–200 (2020)
119. Stoyanovich, J., Howe, B., & Jagadish, H. V.: Responsible data management. Proceedings of the VLDB (Very Large DataBases) Endowment, 13(12), 3474–3488 (2020)
120. Taddeo, M.: Trusting digital technologies correctly. Mind. Mach. **27**(4), 565–568 (2017)
121. Tiddi, I., Schlobach, S.: Knowledge graphs as tools for explainable machine learning: a survey. Artif. Intell. (2022). https://doi.org/10.1016/j.artint.2021.103627
122. Tjong Tjin Tai, T. F. E.: The right to be forgotten: private law enforcement. Int. Rev. Law Comput. Technol. **30**(1-2), 76-83 (2016)
123. Trivedi, S.K.: A study on credit scoring modeling with different feature selection and machine learning approaches. Technol. Soc. (2020). https://doi.org/10.1016/j.techsoc.2020.101413
124. Truby, J., Brown, R., Dahdal, A.: Banking on AI: mandating a proactive approach to AI regulation in the financial sector. Law Financial Mark. Rev. **14**(2), 110–120 (2020)
125. Vakkuri, V., Kemell, K., Kultanen, J., Abrahamsson, P.: The current state of industrial practice in artificial intelligence ethics. IEEE Softw. **37**(4), 50–57 (2020)
126. Verma, M., Kumarguru, P., Deb, S.B., Gupta, A.: Analysing indicator of compromises for ransomware: Leveraging IOCs with machine learning techniques. Proceeding of IEEE International Conference on Intelligence and Security Informatics (ISI) **2018**, 154–159 (2018)
127. von Krogh, G.: Artificial intelligence in organizations: new opportunities for phenomenon-based theorizing. Acad. Manag. Discoveries **4**(4), 404–409 (2018)
128. Wall, L.D.: Some financial regulatory implications of artificial intelligence. J. Econ. Bus. **100**(4), 55–63 (2018)
129. Wang, L.C.: Experience of data analytics in EDA and test: principles, promises, and challenges. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **36**(6), 885–898 (2017)
130. Wee, C.K., Nayak, R.: A novel machine learning approach for database exploitation detection and privilege control. J. Inform. Telecommun. **3**(3), 308–325 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.