【特集】

AIによるESG評価とリスク

越　智　信　仁

要旨

人工知能（AI）によるESG評価は，メディア報道を含む膨大な外部情報源の迅速な解析に強みを有しており，ESGインデックスのばらつきを克服するツールとして，市場規律の向上に資するポテンシャルを有している。他方で，AIのブラックボックス問題や学習データ依存性に起因して，情報源・アルゴリズムの偏向等へのリスクも現状では完全に払拭し切れていない。AIによるESG評価の品質保証は難しい課題であるが，データを中心としたプロセス品質の内部統制に係る検証等に関しては，監査保証論のフレームワークを基礎として役立てられる可能性があり，今後は幅広いAI研究コミュニティと緊密に連携した学際的な研究の深化が求められる。

1　はじめに

これまで数多く設立されてきたESG（環境・社会・ガバナンス）評価・格付機関のESGインデックスは，評価データ・手法等の相違から同一対象に対しても機関によって評価のばらつきが大きいことが実証的に示されており，信頼性のある横比較情報の提供という面では未だ発展途上といえる。こうした中，本稿ではアラベスク社の人口知能（AI）を用いたESG評価ツール（S-Ray）に着目し，ESG評価の比較可能性改善に向けたAIの貢献可能性を指摘したうえで，こうした光の部分（活用可能性）とは裏腹に影の部分（リスク）への対応の必要性を論じる。そこでは，AIのブラックボックス問題や学習データ依存性に起因するESG評価の品質保証に係るリスクの所在とともに，リスクへの対処として，プロセス品質に係る内部統制やデータガバナンスへの信頼性付与の枠組みについても論及する[1)]。

キーワード：人工知能（Artificial Intelligence），ESG評価（ESG Evaluation），比較可能性（Comparability），ブラックボックス問題（Black Box Problem），品質保証（Quality Assurance）

2　ESG評価の比較可能性に向けたAIのポテンシャル

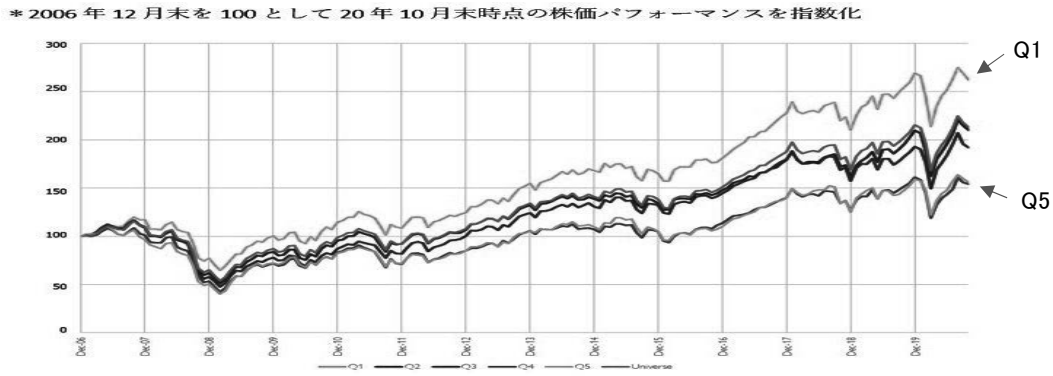
現状の ESG インデックスは，評価データ・手法等の相違から同一対象に対しても機関によって評価のばらつきが大きいことが実証的に示されている（Chatterji et al., 2016; Berg et al., 2022)。このため，企業内での管理情報等の公開の促進とともに，これを受けた評価機関によるデータベースの蓄積・拡充，ESG インデックスの精度向上に向けた評価手法の改善等が課題となっている[2)]。こうした中，従来の評価・格付機関による一般的な ESG 評価手法は，①企業が公表した情報を基に評価する方法，②企業への調査票に対して任意に回答した内容を基に評価する情報が主流であったが，近年では，これに③AI を使った ESG 評価が登場してきている。

AI による ESG 評価機関としては，アラベスク S-Ray 社が世界的に注目を集めているが，もともと同社は，ESG 評価会社の評価に相関性が存在しない課題へのチャレンジにから創業に至った経緯がある。すなわち，評価データのばらつき問題を踏まえ，それらを総合化し比較可能にすることを目的に，英国バークレイズ銀行の社内ベンチャーとしてスタートした後，2013 年にアラベスク・アセット・マネジメント社として分離独立し（創業者セリム），2017 年には AI を用いた Arabesque S-Ray®の開発・リリースに至ったのである。S-Ray は，企業の公開情報と世界 170 か国にわたる 3 万以上の情報源から多くの ESG データを収集・分析しながら，世界約 90 か国，約 9 千社（うち日本企業は約 6 百社）の上場企業に対する日次評価として，アナリストを配置せず人間の主観を排除しつつ，AI による独自のスコアリング・メソドロジーで ESG スコア等を算出しているところに特徴がある。

S-Ray により算出される ESG スコア（0-100 点）は，株価との相関や株価へのインパクトを考慮したマテリアリティ・アプローチを採用しており，膨大な ESG データの中から企業の業績に直結するキーになるものがどれなのかを見つけ出し，ここに AI の機械学習を適用することで独自の分析から相関するファクターを導出する。具体的には，(a) 22 のサステナビリティ課題における企業のパフォーマンスを計測・比較し，(b) 相関性分析（マテリアリティ分析）を通じ，企業の財務パフォーマンスをより良く説明するサステナビリティ課題を特定する。(c) 結果として ESG 投資の際にマテリアリティを考慮することで，どの企業が長期的にアウトパフォームし，投資成果の向上が図れるかを分析することが可能となる。

例えば，雨宮（2020）によれば，ESG 項目を数値化できるようになった 2006 年末から 2020 年 10 月末までの株価と ESG スコアのパフォーマンスについて，S-Ray がカバーする世界企業を ESG スコア順に 5 分割し，最上位 20％を「Q 1」，上位 20％が「Q 2」，中位 20％が「Q 3」，最下位 20％が「Q 5」として比較すると，ESG スコアが高い企業ほど株価パフォーマンスは良好で，同スコアの低い企業の株価パフォーマンスは芳しくない結果が示されている（図 1)。

図 1　世界企業の ESG スコア別株価パフォーマンス比較（時価総額ベース）

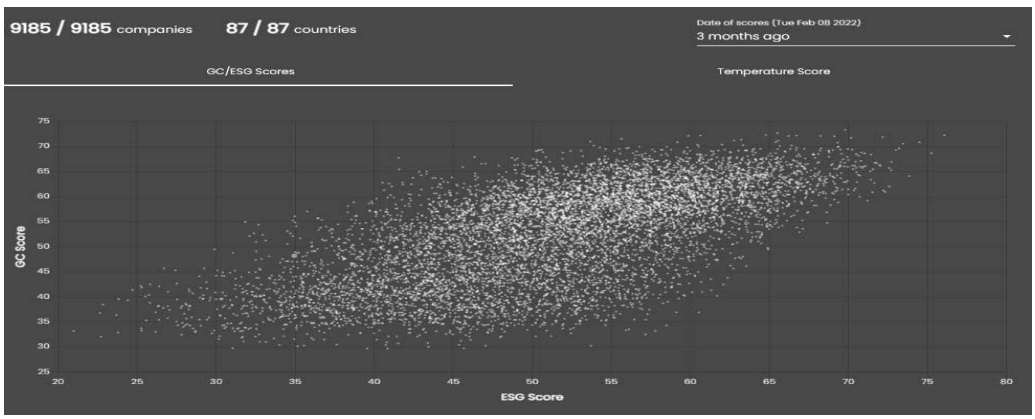


	Universe	Q1	Q2	Q3	Q4	Q5
Return p.a	5.6%	7.2%	5.5%	4.8%	3.2%	3.2%
Volatility p.a	16.4%	15.0%	16.6%	17.4%	18.6%	18.8%
Max Drawdown	-53.7%	-46.0%	-56.1%	-59.2%	-61.1%	-63.5%

（出所）雨宮（2020）

ESG スコアとともに GC スコア（0-100 点）も算出されており，同スコアは国連グローバル・コンパクト原則に基づいて，企業の社会的責任（行動規範）を評価する。あくまで企業の行動自体に焦点が当てられており，ESG スコアと異なり，業界の特異性や財務的重要性を考慮していない規範的な評価である。但し，GC スコアが社会的なインパクトを重視しているといっても，あくまで投資家志向であることに留意が必要である。GC スコアが低い企業は，レピュテーション・リスクが高まることによって，公的なライセンスが失われる可能性などリスク要因で考慮すべき対象となるからである。なお，GC スコアは ESG スコアと緩やかな正の相関関係が窺われる（図 2）。

図 2　世界企業における GC スコアと ESG スコアとの相関



（出所）アラベスク S-Ray 社ホームページ（当日から 3 か月前のデータを一般公表）。

もともと S-Ray というネーミングは，健康診断のレントゲンである X-Ray にちなみ，S は Sustainability ということで，企業の持続可能性の健康診断を可視化する意図が込められている。その際，アラベスク模様が同じパターンが反復する一連の幾何学的美術形式であることにちなみ，日々，約 9 千の企業行動のパターンを比較可能な形で並べて，そこからずれた企業を見出す，すなわち横並びの評価を通じ，あるパターンから逸脱する企業をチェックする。既存の ESG 調査機関が，一つの企業について定性面を含め専門アナリストが細かく分析した企業固有価値の探求に比較優位があるのに対し，S-Ray では個社固有情報の細かな分析が十分できないとしても，競合他社との定量的なスコアで比較可能性を高め投資判断に役立つと同時に，自社の位置付けを横並びで把握可能にするという利用価値がある。

既に，世界の主要な金融機関，機関投資家，企業，コンサルティング会社が S-Ray を使用しており，公表・報道資料等によれば，海外では，スウェーデンの公的年金基金 API，ドイツ銀行，独インデックス開発大手 Solactive，アクセンチュア，BNY メロン，ステートストリート，S&P グローバル，BIMB インベストメント等が S-Ray のデータを使用している。また日本でも，第一生命が S-Ray データによる外国株式 ESG ファンドを組成しているほか，年金積立金管理運用独立行政法人（GPIF）も自らの ESG 評価指標のデータに活用している。こうした AI を活用した ESG 評価への支援機能は，比較可能性の改善を通じて市場規律の向上を促進し，ビジネスと持続可能な社会をつなぐエコシステムの循環に資するポテンシャルを有していると評価できよう（越智, 2021b, 51 頁）。

3　AIによるESG評価のリスク

ESG 評価の分野に限らず，AI の機械学習技術は様々なシステムに埋め込まれ，実社会での応用・実用化が急速に広がっている。しかし，機械学習は決められた範囲内で適切な値を見つけ出すだけで例外に弱く柔軟性がない（松尾，2015，8 頁）中にあって，機械学習技術そのものの研究開発が活発なのに比べて，それらの技術を組み込んだシステムの安全性・信頼性や品質保証に関する研究開発については，まだ取り組みが十分でない。そこにはグローバルに共通の課題があり，わが国でも科学技術振興機構研究開発戦略センターによる CRDS（2018a）（2018b）において，AI ソフトウエア工学の視点から，安全性・信頼性・品質保証に向けた方法論・技術体系の必要性が包括的に論じられている。

深層学習は，人間が形式知化することが難しいレベルの判断規則まで自動獲得し得るということが，人間を上回る精度の達成を可能にしている反面で，機械学習が判定したとき，その理由を人間が理解できないというブラックボックス問題が，システムの安全性を確保するうえで妨げとなっている。すなわち，機械学習技術は，いわば「前例に倣う AI」であり，多数の前例をデータ

として与えると，その中に潜む規則性を見出し，新たに与えられたデータに対して，前例に倣った判断（データの分類・予測・異常検知等）ができるようになるのと裏腹に，人間にとって明示的な規則として理解することが困難なモデル（ブラックボックス）とならざるを得ず，このため出来上がった学習済みモデルにおける深層学習を人間が理解できないので，この解釈性が欠けているという問題が品質管理の課題を生むことになる（CRDS, 2018a, 5 頁; 2018b, 5 頁）。

動作保証・品質保証等の考え方が整備されないまま，機械学習技術を組み込んだシステムの社会実装が急速に進むと，そこで発生した問題や事故が思わぬ社会問題化する懸念が指摘されている（CRDS, 2018b, 8 頁）。まさに両刃の剣であり（Taylor, 2016, p.4），このため基礎研究として，機械学習自体の品質保証，全体システムとしての安全性確保，ブラックボックス問題への対策，問題を効率よく解く工学的な枠組み等が，重要な技術チャレンジとして浮上している。もとのアルゴリズムをきちんと実装できているのかとか，正しさの基準やその検証をどうするか，システム全体の品質をどう評価しどう管理するのか，何らかの問題が起きたときどうするかといった安全性基準の策定（および第三者評価の仕組み）という技術的な課題に加えて，制度・ガイドライン面の整備も急務とされる（CRDS, 2018a，ii , 6 頁; 2018b, 31 頁）。

また，機械学習型システムは，従来のソフトウエア開発と異なり要件の厳密な記述やモジュール毎の検査ができず，学習用データが実データの統計的に適切なサンプルであることは保証されない。こうした中，機械学習自体の品質はデータ集合が左右し，AI は臨機応変ということはできないため，訓練データに潜む差別･偏見をそのまま受け入れてしまう問題や，こうした AI の脆弱性を悪用した攻撃も懸念される（Brundage et al., 2018, p.5）。実際に起きた問題事例として，Microsoft が開発した機械学習型チャットボット Tay が，2016 年 3 月 23 日にインターネット上に公開されて，わずか一日で公開を停止するという事件があった。Tay は，悪意のあるユーザーとの対話から言葉を獲得した結果，差別と陰謀論に染まってしまい，不適切な発言を繰り返すようになってしまったのである（CRDS, 2018b, 8 頁）。

悪意を持った攻撃だけでなく，設計が不十分なため意図しない有害な動作が発生する懸念（Amodei et al., 2016, pp.2-3）のほか，知らずに差別・偏見を取り込んでしまう問題もある。例えば，罪を犯した人間の再犯率を予測するために，人間による判断結果を用いて訓練したとすると，もしその判断を行った人間が特定人種に偏見を持っていたならば，機械学習による判断結果もその傾向を反映してしまう。また，個々の訓練データに偏見を持った判断があるか否かだけでなく，訓練データの分布の偏りや量が差別を生むケースも指摘されている。例えば，特定人種について誤認識が多いことが問題になった事例や，採用判定で性別による差別が起きてしまった事例等が報告されている（CRDS, 2018b, 9 頁）。

AI の社会実装（事業化）においてリスク最小化のマネジメントが重要となることは，ESG 評価においても同様である。そこでは AI による判断や意思決定を行った結果として影響を受ける相手方に対し，アカウンタビリティの履行が求められる（増子，2020，11 頁）。ただ，人間の生

命・身体に実害を与えない ESG 領域では，ミスの深刻性が相対的に低いともいえ，先述した工学的な課題を同様に抱えながらも，実用化サイドにおいて事業展開を比較的早く推し進める可能性がある。その場合には，機械学習型システムの品質管理基準として通常有すべき信頼性を判断するうえで，生命・身体への危害と経済的不利益では，ダブルスタンダードの品質管理基準が首肯され得るか否かも含め，十分に議論を尽くす必要があろう（この点は次節で考察する）。

ESG 評価における AI の利用に関しては，先述したようにスコアリング・モデル（アルゴリズム）がブラックボックス化する中で，情報収集のアルゴリズムがバイアスを含み意図せざる偏向を生む懸念を現状では排除し切れない。また，高いスコアを得ようと近視眼的で形式的な企業行動を助長する場合には，短期的・機会主義的行動によって画一化が横行し，本来的に望ましい ESG 活動の多様性を歪めてしまう懸念もある（水口, 2017; 2018）。また，AI 評価方法自体の信頼性とともに，情報源となる企業開示情報，各種メディアでのニュース報道等に多くを依存することの懸念も残り，チェリー・ピッキングがないとしても，開示企業の ESG／SDGs ウォッシュや特定メディアの偏向報道，誘導記事等を無批判に受け入れて評価を下してしまう危うさを内包していよう（越智, 2019, 26 頁）。

機械は与えられた問題を解くことはできても，問うべき問題を特定することはまだできないので，ブラックボックス型モデルを無批判に受け入れるのではなく，その有効性をクオンツ分析者自身が総合的に判断を下す必要があり（伊藤, 2019, 48 頁），規範的事項のルールメイキングも含め人間が関与するプロセスは不可欠と考えられる。最終的には人間と AI の判断バランスの線引きが重要になり，そこでは人間がＡＩにコントロールされるような社会にしないよう，AI というツールの判断を人間が最終的に確認するとともに，ブラックボックスである AI の判断を用いた責任は誰がどのように負うべきか，AI と人間が共存する社会のありようが問われることになる（瀧, 2020, 2 頁; Soares, 2016, p.6）。

公正なアルゴリズムの利用を技術だけで担保するのは難しく，人間関与原則が AI 社会における公正なアルゴリズムの利用を担保するための基本的考え方となるが，こうした原則的考え方だけで個別具体的な場面の全てにおいて，個人がアルゴリズムによる不公正な評価から逃れることができるわけではない（山本・尾崎, 2018, 102-103 頁）。機械学習のアルゴリズムで示された結果について，相関関係によって導かれた確率的な数字だけしか示されず，開発者を含め誰も理由が説明できない状況を改善すべく，アルゴリズムの透明性確保や説明責任履行に向けて，アルゴリズムやその前提となる使用データに係る信頼性・品質保証に向けた検証のフレームワークが探求されなければならない。

4　信頼性・品質保証に向けた枠組みの構築

4.1　AIガバナンスの重要性とプロセス品質の保証

近年，わが国企業においても AI が意図せず差別などを助長するのを防ぐため，ソニーやパナソニック，日立製作所等では，AI の倫理面での安全性審査，AI の判断基準や製品開発等の際のルール作りなどを通じて，リスク対応を強めているとされる（日本経済新聞, 2021）。こうした動きを後押しし，AI 原則を民間企業・団体が社内規定に具体化していく手助けとして，ガイドラインのような実践的な共通ルール（ソフトロー）を国レベルで策定していくことが望まれる。わが国としての安全性・品質を巡る適切なルールの下で，システム開発のための方法論等が確立されていることは，国際競争力の源泉にもなる（CRDS, 2018b, 12 頁）のであるから，議論の旗振り役ないしファシリテーターとしての国の非拘束的な先導的・啓蒙的関与が求められる。

実際，既に民間事業者等の参考に供するため，AI に関連した契約上の主な課題や論点等を整理した「AI・データの利用に関する契約ガイドライン」（経済産業省, 2018）や，AI 利活用時に留意すべき具体的措置を整理した「AI 利活用ガイドライン」（総務省, 2019）が公表されている。また，各種ガイドラインを相互参照した包括的・実践的手引きとして「AI 原則実践のためのガバナンス・ガイドライン」ver. 1.0（経済産業省, 2021b），ver. 1.1（経済産業省, 2022）も発表されている。こうした非規制的なガイドライン方式は，民間事業者等が自主的なモニタリングや見直しに参照可能な非拘束的な枠組みとして国際的にも有用性が認められており，AI のイノベーションを促進しつつ潜在的ニーズへの対応や機動的なリスク管理に適応し，AI に関係する民間の自主的な合意形成基準の開発や改善に資することが期待される（Executive Office of the President, 2020, pp.6-7）。

検証可能性を含む品質保証の問題は ESG 評価分野に限った話ではないが，AI プロダクトの品質保証に関する調査・体系化，適用支援・ 応用，研究開発を推進するとともに，AI プロダクトの品質に対する適切な理解を啓発する活動として，2018 年に「AI プロダクト品質保証（QA4AI: Quality Assurance for Artificial-Intelligence-based products and services）コンソーシアム」が産学 27 の発起人・団体で設立された。QA4AI は AI プロダクトの品質保証に対する共通の指針として，2019 年には「AI プロダクト品質保証ガイドライン」（2019.05 版）を策定し，その後も随時更新している。そこでは，当該ガイドラインを指針として民間事業者等が，自組織の状況などを熟慮しカスタマイズして反映していくことが想定されている（QA4AI, 2020, 1-4 頁）。

また，国立研究開発法人産業技術総合研究所からは，2020 年 6 月に「機械学習品質マネジメントガイドライン」（第 1 版）が公表されている（2022 年 8 月に第 3 版）。全体として上記 QA4AI ガイドラインが，実際に機械学習 AI を作るエンジニアにとって，機械学習要素の内部品質を改善し作り込むための技術の可能性を見つけ出すための参考書となっているのを受けて，国立研究開発法人のガイドラインは，機械学習 AI を含むシステム全体を企画開発する企業などが，シス

テム全体の利用時品質の確保に向けた必要事項を網羅的に分析・列挙することを意図しており，相互補完関係にあるとされる（産総研, 2021, 115 頁）。

上記の各種ガイドラインにおいて品質保証として本来求められるのは，結果としての「プロダクト品質」であることは言うまでもない。しかし，そのアルゴリズムがオープンにならなければ，学習済みの製品・商品だけを調べても，リバースエンジニアリングで分解したり動作を解析したりして仕様や仕組みを明らかにすることができない以上，学習アルゴリズムを秘匿したままの学習済みの製品・商品の学習結果から，学習アルゴリズムを推定するのは不可能である（松尾, 2015, 246 頁）。学習によって構築されたアルゴリズムは，人間に理解できる状態で変数や重みづけを見出すことは困難（江間, 2019, 36 頁）だからである。

こうした難点への対応として，「説明可能な AI（Explainable Artificial Intelligence ; XAI）」に向けた研究も重ねられている。XAI は，アメリカの国防高等研究計画局（DARPA）が主導しているプロジェクトで，機械学習モデルの出力に加えて，その出力を補助する追加の情報（モデルの解釈，判断根拠の説明など）を出力する技術一般及び研究分野全体を指す用語とされる（原, 2019）。XAI によって説明可能で解釈性の高いモデルが構築されれば，何故 AI がそのような判断を下したのかの要因を説明できるほか，AI が間違った判定を下した場合に何故間違ったのかも検証可能となり，AI による判断というプロダクト品質の品質保証や信頼性付与に向けても大きな貢献が可能であろう[3)]。

また，XAI によりモデル予測の特徴量の貢献度が分かれば，人間が解釈しやすい注釈を加えるなどして，例えば医療診断・金融取引等の分野においても，エンドユーザーとの信頼関係構築や透明性の向上が図れる。同様に，ESG 評価を含む市場運用の場においても，AI のブラックボックス化の影響により，投資家による市場ダイナミクスの変化の解釈や新たに出現するリスクの特定や理解をより困難にする可能性があるとすれば，XAI は，長期的に見て金融システム全体の防壁の維持に役立つ可能性がある（Deloitte, 2020, 9, 11 頁）。

しかし，学習アルゴリズムの内実を完全に捕捉し切れない（完全なホワイトボックス型 XAI を実現できない）段階では，次善の策として，プロダクト品質の測り方を「プロセス品質」の保証に切り替える必要もあるかもしれない。ソフトウエア工学の視点からの AI 品質保証（安全性等）の分析は筆者の能力を超え監査保証論の射程からも外れるが，モデル策定プロセスの内部統制に係る検証等に関しては，監査保証論のフレームワークが基礎として役立てられる可能性がある。AI による評価一般の信頼性確保に向けた考察は引き続き難しい今後の課題ながら，AI による ESG 評価という非財務情報の策定プロセス（内部統制）の適切な構築と運用については，監査保証の実務や研究が貢献できる余地はあるのではなかろうか。

AI による ESG 評価情報の信頼性をより広い文脈で捉えた場合には，AI 利用側のガバナンス，プロセスに係る内部統制（とりわけ内部監査）による信頼性確保が先決的に重要となる。AI 利用に係る企業組織の自律・開示の仕組みの実効性ある整備が求められ，こうした思考には統合報告

など非財務情報の作成・開示における信頼性付与とも類似した側面がある。例えば統合報告書では，ガバナンスに責任を負う者が情報開示の誠実性を確保する責任を有しているという認識等の表明を求めている（IIRC, 2013, 1.20項）。また，南アフリカのキングⅢ（IoDSA, 2009）で提唱されたCombined Assuranceにおいては，外部の保証だけに焦点を当てるのではなく，経営者等による実効的な枠組み整備・確認とともに，内部監査等の保証機能も重視している。

同様の発想は古くからみられており，2003年に英国勅許会計士協会（ICAEW）が，作成プロセスに着目したOFR（Operating and Financial Review）の保証という基本思考の下で，OFR作成プロセスに係る中間指針（ICAEW, 2003）を公表した際にも窺われた。その中で取締役のためのOFRの六つの作成原則を明示しており，OFRは組織化された作成プロセスから創出され，取締役会に対する日常の報告と承認を基礎とすべきであって，企業組織における内部報告とOFRの内容について整合性を求めていた。そこでは，第三者が作成プロセス自体の信頼性を確保する前提として，あくまでも企業の自律的な財務報告プロセスを基礎とすることが重要と考えられたのである（古庄, 2011, 46頁）。

このように非財務情報の作成への信頼性付与におけると同じ観点から，AIを利用したESG評価の場合にも，基礎となる情報の作成プロセスへの社内外連携によるガバナンス・監視体制の強化・高度化が求められる。AIガバナンスの実践的ガイドライン（経済産業省, 2021b）では，AIシステム開発企業はAIガバナンスの実践状況を，コーポレートガバナンス・コードの非財務情報に位置付け積極的な開示を検討すべきとしたうえで，AI倫理の取り組みを統合報告書に記載することを含め，経営層からの意欲的な情報発信を検討している企業実践事例を紹介している（43-44頁）。また，システムの設計・運用から独立した者に，システムの設計や運用を含むAIマネジメントシステムが適切に機能しているか検証すべきとして，内部監査部門等を活用している企業実践事例にも言及している（45-47頁）。

こうした考え方はOECDのAI原則にもみられており，包摂的な成長，持続可能な開発及び幸福，人間中心の価値観及び公平性，セキュリティ及び安全性と並んで，透明性及び説明可能性や頑健性，アカウンタビリティが挙げられている（OECD, 2019, pp.7-8）。わが国でも「AI利活用原則案」（総務省, 2018）や「人間中心のAI原則」（内閣府, 2019）が策定されているが，こうしたAIに関する原則については，概ね国際的にコンセンサスが形成されてきている（総務省, 2019, 2頁）。実際，世界の産官学等におけるAI原則を分類した研究（Zeng et al., 2019; Fjeld et al., 2020）によれば，各国の原則は，透明性と説明可能性のほか，プライバシー，アカウンタビリティ，安全性とセキュリティ，公正性と非差別性，人間による技術管理，専門家の責任，人間的な価値の促進などに収れんするとされる。

4.2　リスクベースのデータガバナンスと多層的信頼性付与

社内ガバナンス・内部統制の確保を前提に，外部からの規制的なAIガバナンスの付加を議論

していくに際しては，規制の程度を AI 利活用のリスクの大きさに対応させる考え方（リスクベース・アプローチ）が国際的なコンセンサスとなっている（経済産業省, 2021a, 6 頁）。例えば，欧州委員会の AI 規制法案（EC, 2021）では，３分類した AI リスク（容認できないリスク，高リスク，低・最低限のリスク）に相応しい規制等を導入する方向性を示している。また，英国のデジタル・文化・メディア・スポーツ省に設置されたデータ倫理・イノベーションセンターでも，AI 利用の具体的な応用類型別にリスクを検討した調査報告書の中で，各種リスクを高・中・低の３段階で評価していた（UK, 2020, 8.Summary of Findings- Risks）。

リスクが相対的に高い分野では規制的な必要性が高まるであろうし，リスクが低ければより，より緩やかな自主規制や開示規律等の手法も取り入れる余地が広がる。一口に AI 応用といっても，例えば自動運転と医療機器では AI 応用によって潜在的なリスクは異なるのであるから，業界の事情を考慮しながら，幾つかの規制的・自主的手法を組み合わせて，リスクに応じて柔軟に対応する必要がある。また，小規模な AI 利活用におけるリスクへの対応や組織のあり方は，大規模な利活用におけるそれとは当然異なるであろうし，具体的な AI の利活用の場面によってもリスクは異なる。この点に関し，AI を活用した ESG 評価情報の経済取引での利活用に限れば，リスク度という点で医療等の人命・身体の安全に直結する分野と同列には考えられないので，規制的・拘束的なルール化の必要性は乏しいかもしれない。

一般に品質保証とは，あらゆる条件における動作の正当性を将来にわたり予測することによって達成するものであるが（QA4AI, 2020, 1-2 頁），AI の利活用に伴うリスクないし負のインパクトは，用途，範囲，使用態様等によって大きく異なり，AI の ESG 評価における信頼性・品質保証において最重要なのは，各種品質保証ガイドラインで取り上げられている項目の中で公平性の問題，すなわち情報利用者に影響するバイアス除去（不公平な判断の回避）への取り組みではなかろうか。先述した QA4AI の「AI プロダクト品質保証ガイドライン」では，AI プロダクトの品質保証において考慮すべき軸として，「Data Integrity」「Model Robustness」「System Quality」「Process Agility」「Customer Expectation」の 5 つの軸を提案しているが，AI による ESG 評価に係るプロダクト開発においては，データ元の偏りや不公平な判断につながりかねない特徴量の識別など，データの質（整合性）が相対的に重要になると考えられる。

AI による ESG 評価情報の利用者は，AI によりナッジ（合理的選択のための支援）を行う場合が多く，AI 利活用における「消費者的利用者」に該当し，その意思決定・感情が誤導されるリスクをミニマイズするガバナンスが求められる。AI システムの用途やリスクレベルに応じて，使用するデータの来歴，品質に求められる管理の仕組みは異なり，田丸ほか（2020）では政府が AI の消費的利用者になる場合に，リスク度に応じて考慮すべき AI の学習データに関する透明性の確保，偏り（バイアス）の排除，データ加工の来歴保管の必要性などを論じている。ESG 評価に用いるデータセットに関しても，その精度やきめ細かさが最終的な AI システムのパフォーマンスに大きく影響するのであるから，AI の学習等に用いられるデータの代表性やデータに内在す

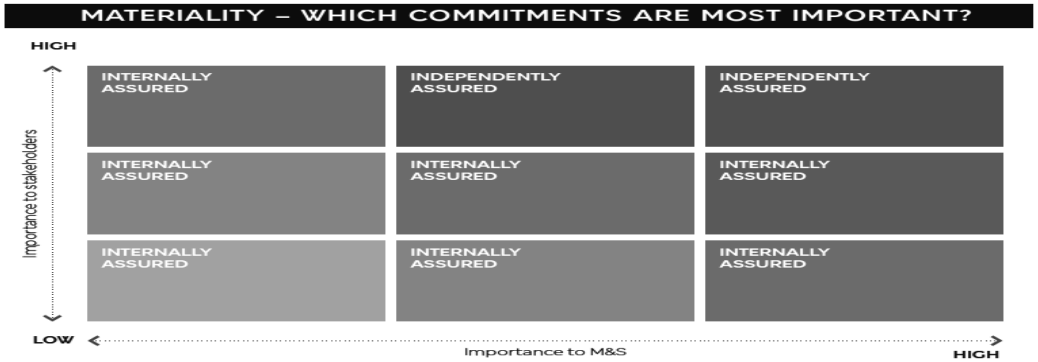
る社会的なバイアスといったデータの質に留意する仕組みが求められる。

使用データを含む AI の信頼性確保の枠組み構築には工学的アプローチのみならず，企業スタンスの明確化や社会との対話など広範におよぶ様々な取り組みが必要となる。サービスの提供者・AI モデルの開発者・データの提供者等において，公平性を維持する具体的な仕組みとして，開発・運用プロセスの明確化，AI の公平性に係るリスク要因・データのバイアス除去に向け，データの活用を進めるデータ・ポリシーの確立，AI 活用時の倫理教育や運用に関する社内規定の整備・専門人材の確保，リスクアペタイトの設定など，組織レベルでのガバナンスが重要となる（経団連, 2019, 36, 43 頁）。大量のデータを処理する AI において，すべてのデータの特性を利用者に分かるように「説明」することは困難だとしても，AI に用いるデータセットに関して，利用目的や社会的な要請を踏まえ学習データセットの品質・透明性確保に向け，不正・偏向データやノイズの除去，データの確からしさや来歴・出典等の情報を明らかにできるようにすることが重要である（同 37 頁）。

AI システムの開発側にはモデル・データに係るガバナンスの説明責任が求められる一方，AI による ESG 評価への信頼性付与に際しては，情報利用者が，信頼された情報をみて，どのような経済的意思決定をするか，その意思決定が信頼性を付与することによってどれだけ有用性が増すかとのベネフィットや，信頼性付与のためのコストを踏まえ，全体としてのコスト・ベネフィット判断の下で，保証を含む信頼性付与の程度（投資家にどういう信頼性付与を行うことにニーズがあるのか）が検討されなければならない。そこでは，AI 利活用情報の重要性に関するグラデーションやコスト・ベネフィットを総合的に勘案し，広い意味での信用を事実上支援する手段について，マトリックスでの検討を深める必要があろう。

例えば，AI を用いた情報ではないが ESG 開示の信頼性付与に関し英国 Marks & Spencer 社の年次報告書では，非財務情報の重要性に応じて外部保証と内部監査を使い分けている（図 3）。同社は，戦略とビジネスモデルの中核となる「環境，人権，社会貢献などの包括的なプログラム」を「プラン A」と名づけ，その詳細を戦略報告書とは別に「プラン A 報告書」として説明しており，2018 年度までの同報告書の中で，重要性マトッリクスに応じた信頼性付与の対応が記載されていた。そこでは，重要性マトリックスにおいて，横軸（M&S 社にとっての重要性 3 区分）と縦軸（ステークホルダーにとっての重要性 3 区分）に応じて 9 区画を識別し，最も重要な 2 つの区画は外部保証を割り当てる一方，その他は内部監査により対応している。同様に，AI を用いた情報についても，その重要性やコスト・ベネフィット等に応じて，信頼性付与の手段も多様であり得よう。

図 3　英 M&S 社の重要性に応じた信頼性付与



（出所）M&S, 2018, p.36

なお，IAASB による非財務情報（EER：Extended External Reporting）保証に向けた規範性のないガイダンス文書の関連サポート資料（IAASB,2021, p.2）には，EER の信憑性（credibility）に関し，監査保証以外にも，健全な EER フレームワーク，報告プロセスに対する強力なガバナンス，一貫性のある幅広い情報などによっても高められると指摘している。同様に AI による ESG 評価情報についても，様々な信憑性向上策を活用する余地があり，リスクベース・アプローチを適用する中で，AI 利活用情報の重要性に関するグラデーションやコスト・ベネフィットを総合的に勘案し，広い意味での信憑性を事実上支援する手段について，マトリックスでの検討を深める必要があろう。

そこでは経済産業省（2021a）44-45 頁でも指摘しているように，一定の重要なリスク事項についてハードローないしガイドラインによる開示や，外部監査人あるいは第三者による内部統制監査（合理的保証）・レビュー（限定的保証）といった保証（assurance）を求めることも考えられる。さらに信頼性（reliability）付与の意見表明業務（保証）の範疇からは外れるが情報利用者の信憑性（credibility）を高める方策として，業務プロセスの自主的改変や内部監査の強化のほか，限定的保証よりもコストが相対的に低い合意された手続（agreed upon procedures）など非保証業務の支援可能性，AI 専門家による意見付与サービスの活用可能性も俎上に載るであろう。

5　おわりに

AI の活用は，前例のない社会課題に取り組むために重要な役割を果たし，民間企業における新規ビジネス創出に向けた用途が広がっている（Deloitte，2021，6 頁）。そうした中で，経済発展とともに SDGs（Sustainable Development Goals）を達成するための鍵となり得る（内閣府，

2019, 1 頁; EC, 2020, p.2）反面で，予測困難性等に起因して市民社会への重大なリスクも秘めている。AI の機械学習は静的に記述されたルールではなく，データを用いた学習過程に依存した統計的な処理によって，変数の重みが動的に調整され出力結果が逐次変更（改善）されていくため，最終的出力結果に至る導出過程が良くも悪くも人間の直感的もしくは論理的解釈を超えたものになってきている。

解釈可能性に欠ける状況下では，AI を構成するプログラムの内容を対象に，従来のソフトウエアの品質保証の手法によって審査することは難しく，どのような要素が影響して一定の判断が下されたかを事後的に検証することは開発者といえども無理である（経済産業省, 2020, 22-23 頁）。こうした中，システムの設計・運用におけるアカウンタビリティ，あるいは開発主体側の社会からの信頼確保に向けたコーポレート・ガバナンスとして，AI システム（アルゴリズム）のプロダクト品質を代替するプロセス品質の信頼性確保に向け，ガバナンス原則・ガイダンスに沿った開発倫理基準/行動規範等に適合的なリスクの把握・評価など，プロセス品質に係る内部統制を利害関係者に説明（開示）することが，まず求められる。

次いで，説明（開示）情報の検証を考えるに際しては，経済取引に資する（生命・身体の安全性に影響しない）AI の ESG 評価分野では，リスクベース・アプローチを基軸に規制的・拘束的ではない管理手法が適当であろう。そこでは，情報利用者の公平性・公正性を阻害するリスク・重要度に応じ，コスト・ベネフィットのマトリックスの下，開発・運用プロセスに係る内部統制のほか，アルゴリズムの前提となるデータの選別・保存・加工・共有・追跡可能性等に関し，監査保証による第三者意見表明やそれ以外の信憑性付与手段（内部監査，合意された手続等）を，多段階で割り当てる方向性が考えられる。AI による ESG 評価の信頼性付与に向けた具体策については，会計・監査保証領域においても，幅広い AI 研究コミュニティと緊密に連携した学際的な研究の深化が今後とも求められる。

注

1）本稿は，越智（2021a）（2021b）を基礎とした 2022 年度日本社会関連会計学会統一論題報告に，その後の状況や越智（2022）等を踏まえて加筆・修正したものである。

2）ESG 評価における規範的判断の独自性・多様性は首肯されるとして，ばらつき問題で俎上に載せているのは，むしろ評価の未成熟（データの制約・偏り，過渡的な評価手法等）に起因するノイズが少なくないのではないかという点であり，評価の多様性と未成熟は峻別して議論する必要があろう（越智, 2023, 8 頁）。

3）但し，そこで説明可能なのは予測の根拠（モデルが重要視した部分），すなわちどの特徴量（変数）がモデルの予測結果にどれだけ貢献しているか（値の正負，大小等）という特徴量の識別に止まり，何故そこを重要視したのか，どのような計算をして重要視したのかなど，AI モデルの中身を明らかにするものではない（大坪ほか, 2021, 231 頁; 宇佐美, 2020）。XAI によりモデル予測に与える重要な特徴量を知るだけでも品質保証に向けて大きな前進ではあるが，モデルの計算過程（モデルのふるまいを表す論理的根拠）には未だ別の意味でブラックボックス性を残しており（村田, 2020, 51 頁; 宇佐見, 2020），特徴量の相関関係だけでなく因果関係にまで言及できる XAI は限られる（大坪ほか, 2021, 65 頁）のが現状とされる。

参考文献

Amodei, Dario., Chris Olah, Jacob Steinhardt, Paul Christiano,John Schulman, and Dan Mané (2016) *Concrete Problems in AI Safety*, arXiv:1606.06565 [cs.AI].

Berg, F., Kölbel, J. F., and Rigobon, R. (2022) Aggregate Confusion: The Divergence of ESG Ratings, *Review of Finance* 26 (6), pp.1315-1344, https://doi.org/10.1093/rof/rfac033

Brundage, Miles et al. (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, arXiv:1802.07228 [cs.AI].

EC: European Commission (2020) *White Paper on Artificial Intelligence – A European approach to excellence and trust* (February 19).

EC (2021) *Proposal for a Regulation of the European Parliament and the Council laying down requirements for Artificial Intelligence* (April 21).

Executive Office of the President (2020) *Office of Management and Budget, Memorandum to the Heads of Executive Departments and Agencies, Guidance for Regulation of Artificial Intelligence Applications* (November 17).

Fjeld, Jessica., Nele Achten, Hannah Hilligoss, Adam Christopher Nagy, Madhulika Srikumar (2020) *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rightsbased Approaches to Principles for AI,* Berkman Klein Center for Internet & Society.

IAASB: International Auditing and Assurance Standards Board (2021) *Credibility and Trust Model relating to Extended External Reporting (EER), in Non-Authoritative Guidance on Applying ISAE 3000 (Revised) to Extended External Reporting (EER) Assurance Engagements*.

ICAEW: Institute of Chartered Accountants in England and Wales (2003) *Preparing an Operating and Financial Review: Interim Process Guidance for UK Directors.*

IoDSA: Institute of Directors of South Africa (2009) *King Code of Governance for South Africa 2009*, http://www.ecgi.org/codes/documents/king3.pdf

IIRC: International Integrated Reporting Council (2013) *The International <IR> Framework*.

M&S: Marks & Spencer (2018) Transformation Underway: Plan A Report 2018.

OECD: Organisation for Economic Co-operation and Development (2019) *Recommendation of the Council on Artificial Intelligence.*
https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Soares, Nate (2016)The Value Learning Problem, In: Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence,
https://intelligence.org/files/ValueLearningProblem.pdf

Taylor, Jessica., Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch (2016) *Alignment for Advanced Machine Learning Systems*, MIRI,
https://intelligence.org/files/AlignmentMachineLearning.pdf

UK: United Kingdom (2020) *AI Barometer, Center for Data Ethics and Innovation*,
https://www.gov.uk/government/publications/cdei-ai-barometer.

伊藤敬介（2019）「AI，ビッグデータとクオンツ運用」『証券アナリストジャーナル』第 57 巻第 4 号，44-49 頁。

宇佐見一平（2020）「説明可能な AI とは」Ledge. ai 株式会社 2 月 12 日。

江間有紗（2019）『AI 社会の歩き方―人工知能とどう向き合うか』化学同人。

大坪直樹・中江俊博・深沢祐太・豊岡祥・坂元哲平・佐藤誠・五十嵐健太・市原大暉・堀内新吾（2021）『XAI（説明可能な AI）―そのとき人工知能はどう考えたのか？』リックテレコム。
越智信仁（2019）「比較可能な ESG 評価の可能性と課題」『インベスター・リレーションズ』第 13 巻第 1 号，17-31 頁。
越智信仁（2021a）「ESG エコシステムに向けた AI の貢献可能性」『総合政策研究紀要』第 37 号，33-53 頁。
越智信仁（2021b）「ESG 評価における AI の活用事例」日本会計研究学会スタディグループ「AI 技術の活用による ESG 情報の評価に関する研究」（中間報告書）第 4 章所収，41-54 頁。
越智信仁（2022）「AI による ESG 評価の信頼性・品質保証」日本会計研究学会スタディグループ「AI 技術の活用による ESG 情報の評価に関する研究」（最終報告書）第 5 章所収，60-79 頁。
越智信仁（2023）「ESG 評価情報の意義と課題―行動規範の実践へ」『証券アナリストジャーナル』第 61 巻第 2 号，6-14 頁。
QA4AI: AI プロダクト品質保証コンソーシアム（2020）「AI プロダクト品質保証ガイドライン」（2020.08 版）。
経済産業省（2018）「AI・データの利用に関する契約ガイドライン」。
経済産業省（2020）「Governance Innovation: Society 5.0 の実現に向けた法とアーキテクチャーのリ・デザイン」。
経済産業省（2021a）「我が国の AI ガバナンスの在り方 ver. 1.0（AI 社会実装アーキテクチャー検討会 中間報告書）」。
経済産業省（2021b）「AI 原則実践のためのガバナンス・ガイドライン ver. 1.0」。
経済産業省（2022）「AI 原則実践のためのガバナンス・ガイドライン ver. 1.1」。
経団連：日本経済団体連合会（2019）「AI 活用戦略―AI-Ready 社会の実現に向けて」。
産総研：国立研究開発法人産業技術総合研究所（2021）「機械学習品質マネジメントガイドライン」サイバーフィジカルセキュリティ研究センター・テクニカルレポート，CPSEC-TR-2020001。
CRDS：国立研究開発法人科学技術振興機構（JST）研究開発戦略センター（2018a）「俯瞰ワークショップ報告書　機械学習型システム開発のパラダイム転換」，CRDS-FY2017-WR-11。
CRDS（2018b）「（戦略プロポーザル）AI 応用システムの安全性・信頼性を確保する新世代ソフトウエア工学の確立」，CRDS-FY2018-SP-03。
総務省（2018）「AI 利活用原則案」（情報通信政策研究所）。
総務省（2019）「AI 利活用ガイドライン―AI 利活用のためのプラクティカルリファレンス」（AI ネットワーク社会推進会議）。
瀧博（2020）「序論―テクノロジーの進化と監査における研究課題」瀧博編著『テクノロジーの進化と監査―AI とデジタル技術が拓く新たな監査の可能性』同文舘出版，1-9 頁。
田丸健三郎・満塩尚史・柳原尚史・高木幸一・西村毅・梅谷晃宏・楠正憲・細川努（2020）「AI システムにおけるデータ利用の特性と取扱い上の留意点」，政府 CIO 補佐官等ディスカッションペーパー。
Deloitte（2020）「未知の領域の舵取り―金融サービスにおける人工知能の戦略上・規制上のリスクを理解する」。
Deloitte（2021）「AI ガバナンスサーベイ―社会情勢の変化に伴う AI 利活用と投資状況の変容」。
内閣府（2019）「人間中心の AI 原則」（統合イノベーション戦略推進会議決定）。
日本経済新聞（2021）「パナソニックも AI 指針」5 月 17 日朝刊。
原聡（2019）「私のブックマーク 説明可能 AI（Explainable AI）」『人工知能』第 34 巻第 4 号，577-582 頁。
古庄修（2011）「統合的財務報告におけるナラティブ情報の信頼性確保と保証問題」『産業経営研究』第 33 号，39-50 頁。

増子和起（2020）「AI の利用とアカウンタビリティ」日本社会関連会計学会スタディグループ最終報告書『SDGs 時代の社会的評価指標開発への挑戦』，8-12 頁。
松尾豊（2015）『人工知能は人間を超えるか―ディープラーニングの先にあるもの』KADOKAWA。
水口剛（2017）「AI と ESG 評価」QuickESG 研究所 6 月 12 日。
水口剛（2018）「テクノロジー産業と ESG 投資―その光と影」QuickESG 研究所 12 月 11 日。
村田潔（2020）「説明可能な AI（XAI）」日本情報経営学会第 80 回全国大会予稿集，51-54 頁。
山本龍彦・尾崎愛美（2018）「アルゴリズムと公正」『科学技術社会論研究』第 16 号，96-107 頁。

（付記）本稿は JSPS 科研費 No.22H00897 の助成を受けた研究成果の一部である。

越智　信仁（関東学院大学経営学部）