連載 プロマネの現場から
第 164 回 AI の課題と「責任ある AI」

蒼海憲治（大手 SI 企業・・製造業系事業部門・技術総括部長）

AI は社会・業界を変革し、実世界の重要な問題を大規模に解決しています。しかし、AI による問題解決が新たな問題を生じさせないためには、すべての人にとって役立つ AI を構築するという大きな責任が伴います。

AI 技術に関するさまざまな側面を学習する中で、気になったことは「責任ある AI」という考え方・取り組みでした。「責任ある AI」とは、従業員や企業に力を与え、顧客や社会に公正な影響を与えるために、善意を持って AI を設計、開発、展開することであり、企業が信頼を得て自信を持って AI をビジネス活用できるようにすることを指します。

一般的に、人間による意思決定にはバイアスが入りますが、AI の判断は客観的・中立であると思う傾向があると思います。

しかし、実際にはそうなりません。たとえ AI の計算エンジンそのものが中立なものであったとしても、AI にインプットされたデータそのものに、人間によるバイアスが入っている場合、AI からのアウトプットもバイアスが入った結果になってしまうからです。まさに、「ゴミをつめれば、ゴミを吐き出す（Garbage in, garbage out.）」というコンピュータ工学の格言は、AI によるシステムにおいても同様に適用されます。

AI の問題事例、AI による「暴言」「人種差別」「男女差別」などについては、たびたびニュースになるため、ご存じの方も多いかと思いますが、AI の「暴走」を防ぐことを考えるにあたって、頻発する AI トラブルの例を振り返ってみます。

・2015 年には、グーグルフォトが黒人系カップルの写真を「ゴリラ」と自動的にタグ付け。写真に写っていた男性のツイートにより、問題が発覚しました。グーグルは問題があったことを認め、すべての「ゴリラ」のタグの削除や、「ゴリラ」のキーワードでの検索停止などの対応策を表明しました。この事例は、元はといえば AI が学習していた「黒人の画像データ」が「白人の画像データ」と比べると少なく、それが偏見のある結果につながったのだろうと推測されています。

・2016 年、マイクロソフトは、自分で会話できる AI ボット・Tay が暴言を吐くようになってしまったため、サービスを停止。

・2018 年、アマゾンの顔認識 AI「レコグニション」がアメリカの上下両院 28 人の議員の顔を「犯罪者」だと誤認識しました。問題なのは、黒人系の議員の誤認識率が高かったという点です。

・2018 年、マサチューセッツ工科大学の研究者が大手ベンダーの顔認識技術を調査した結果、男性より女性、白人より黒人の顔認識の精度が低いと公表。

・2018 年、フェイスブックが収集するユーザデータの一部が、イギリスの政治コンサルティング企業・ケンブリッジ・アナリテイカ社にわたり政治利用される。

・2018 年、ウーバーの自動運転車が死亡事故を起こす。

・2019 年、ゴールドマン・サックスは、Apple Card の利用者の信用スコアを算出する際、女性に不当に低いスコアが付けられ、クレジットカード限度額に差が生じていたことに対し、非難を受ける。

・2019 年、フェイスブックが広告主の指定したターゲットとは関係なく、雇用や住宅差別を固定化する広告推奨アルゴリズムを使っていたことが判明する。

・2020 年、白人警察官による黒人死亡事件に端を発した「ブラック・ライブズ・マター」の中、顔認識 AI の誤認識による危険性に注目が集まる。

・2021 年、韓国のスタートアップ、スキャッター・ラボが「寂しさを和らげる友達」をコンセプトに、差別語を除外する仕組みを取り入れて開発された「イ・ルダ」という名の会話型 AI が差別的発言をしていることがネットで話題になり、サービス停止となる。

他にも、IBM やマイクロソフトの顔分析技術は白人男性に関しては非常に精度が高い一方で、有色人種や女性に対してはエラーが起こりやすいことが明らかになりました。その後、2020 年にいずれの企業も顔認識 AI からの（一時的なものも含む）撤退を表明しています。

そして同じく 2018 年、アマゾンが AI を用いた人材採用ツールの開発を進めていたところ、技術職では男性ばかりを採用しようとする傾向があることがわかりました。AI が過去の採用履歴や履歴書のプロフィールを学習した結果、「この会社では男性を採用することが好ましいと」判断したといいます。結局、これらのプロジェクトチームも差別を助長しかねないとして解散に至っています。

実際、AI によって消費者が被る被害には、さきほどの事例のように多岐にわたります。

①社会的有害性：固定概念が AI により増強されることによる被害
②経済損失　　：特定のグループに属する人たちの経済的な選択肢が狭まる
③機会損失　　：属するグループによって仕事・住宅・教育などへのアクセスに差が生じる
④自由の損失　：監視、人権、言論の自由などの制約

そのため、AI に対する問題意識の高まりの中、世界各国・地域で法規制の動きも加速しています。

しかし、法規制よりも、技術の進展や社会への適用の方が先に進む中、AI の倫理的なガバナンスの制定が企業や組織に求められるようになっています。VUCA 時代において、法律を順守するだけでは十分ではなく、法律が規定されていない場合でも、強い倫理観を持つことが大切である、との指摘が、「責任ある AI」のベースになっています。

以上のような背景を踏まえて、「責任ある AI」に対する取り組みは、AI 技術を扱う企業・組織にとって必須となりつつあります。各社における「責任ある AI」の定義はこうです。

グーグルは、≪責任範囲が明確な AI を構築する方法の厳格な評価は正しい手法であるだけでなく、成功する AI を作成するために不可欠な要素であると確信しています。≫

マイクロソフトでは≪責任ある AI における 6 つの基本原則として、アカウンタビリティ、包括性、信頼性と安全性、公平性、透明性、プライバシーとセキュリティを挙げています。これらの原則は、AI がより主流な製品やサービスに活用されていく上で、責任があり、信頼できる AI を生み出すために不可欠です。≫

アクセンチュアによると≪「責任ある AI（レスポンシブル AI)」とは、顧客や社会に対して AI の公平性・透明性を担保する方法論です。これに基づいて AI を設計・構築・展開することで、真に人間中心の AI 活用を実現します。≫

では、求められる「責任ある AI」とはどのようなものでしょうか。

ここでは、アクセンチュアの「責任ある AI」の概念を、『責任ある AI―「AI 倫理」戦略ハンドブック』（＊）から紹介したいと思います。

「責任ある AI」の実践にあたって、５つの行動原則（ＴＲＵＳＴ）を規定しています。

Ｔ：信用できる（Trustworthy）
AI の設計・構築時、安全性を重視し、物事に誠実に向き合い、多様で広い視点を持つ、という実績を一つ一つ積み上げる。

Ｒ：信頼できる（Reliable）
積み上げられた信用から、将来の高度な判断とより良い意思決定への支持を集める。

Ｕ：理解できる（Understandable）
信用を得るためには、AI が透明性を持ち、人によって解釈可能である必要がある。

Ｓ：安全が保たれている（Secure）
信用を得るためには、企業や顧客の情報・データのプライバシーに配慮し、安全性を確保しなければならない。

Ｔ：共に学びあう（Teachable）
このような信用・信頼を勝ち得た AI と人間とが情報交換し、共創し、相互の情報提供、相互教育をする世界を実現する、人間中心のデザインを目指していく。

これら「ＴＲＵＳＴ」の中心に、「AI 倫理」が位置づけられています。

次に、AI の潜在的リスクに備える４つの観点からアプローチしています。
AI は潜在的なリスクをはらみやすい技術であり、主に「目的と影響」「入出力」「アルゴリズム（計算方法)」「データ」でリスクが混入しやすくなっています。
AI はデータに存在するパターンを学習し、未知のデータが入力された際に、そのパターンに反応して結果を出すものある。したがって、AI を開発・利活用する際には、以下の点に留意しなければなりません。

・どのような目的で AI を活用するのか、AI がどういった影響を及ぼすか。
・AI に対して、どのような入力と出力を想定するのか。
・AI をどのようなアルゴリズム（計算方法）で実装するか。
・AI の学習のためにどのようなデータを使うか。

（１）目的と影響：

例として、AI による情報収集の効率化のため、大量のニュースや情報から、自分の関心やプロフィールに合わせて記事を収集してくれるキュレーションシステムを考えます。個人の趣味嗜好・関心に合わせて情報を集めるという目的には一見問題がないように思えます。しかし、実際には、キュレーションシステムにより提示される情報は、偏ってしまうというリスクがあります。

「エコーチェンバー」と呼ばれる、閉鎖的な空間の中で特定の意見・信念が増強されてしまい、それ以外の意見が受け入れられなくなってしまう現象です。

２０２１年１月６日の米国連邦議会議事堂占拠事件の背景には、「エコーチェンバー」により増強された陰謀論があったとされています。

（2）入出力：

AI への意図していない入力、AI からの意図していない出力により、これまで想定していなかった結果を生む可能性があります。

（3）アルゴリズム（計算方法）：

「アルゴリズムとはプログラムに埋め込まれた意見である」（キャシー・オニール）という指摘もあります。

その理由は、設計者が何を成功とするかという基準を決め、その基準に基づいてアルゴリズムが設計されるためです。人間が設計するものである以上、設計者が望みも気づきもしない形でバイアスの影響を受けています。しかもより問題なのは、アルゴリズムは客観的なものであり、その出力は合理的であると判断してしまう傾向があるので、そのバイアスに気づくことができないケースの方が多い点にあります。

人間である AI 設計者には「確証バイアス」があるため、「チェリーピッキング」を行ってしまいます。つまり、アルゴリズムの設計段階において、どういう特徴量を「使って」、どういう特徴量を「使わない」のかを、自分の仮説を強化するように選択して設計してしまいます。

（4）データ：

使っているデータそのものに問題があるケースです。大規模・大量のデータの中には、偏見・差別的な表現が含まれおり、また、データの収集方法によっては内容に偏りが生じています。こうしたデータに基づいて学習された AI システムは、当然こういった偏見・差別表現を内包したものになっています。

また、AI における重要な学習方法の一つは「教師あり学習」、すなわち、システムとして出力してほしい解（教師）と入力をペアで与えることで、理想的な解を出力できるよう

に AI を訓練していく方法です。解を与えるのは人間の作業になりますが、以前から、この人間によって与えられる解にバイアスが混入することが指摘されています。

そのため、AI を開発するための必須のデータを、誰が作成したのか、どのように収集したのか、バイアスが含まれていないかどうかの公平性・透明性が強く求められています。

ここで、重要なポイントは、AI の説明可能性と解釈可能性になります。
説明可能性とは、AI 内部のアルゴリズム等を考慮した上で、入力と出力との関係を説明できるということです。これは、万が一、AI の判断ミスが生じた際、その原因究明ができるということになります。
解釈可能性とは、内部の仕組みを知らなくても、入力に対する出力を予測できるということです。
したがって、現場を知らない AI 開発者が、データだけを使って開発した AI を現場に投入することはとても危険なものになります。

≪データの収集から AI のアルゴリズム設計、学習と結果の評価のすべての段階で現場のベテラン、管理者、新人などあらゆる関係者を積極的に巻き込み、開発を進めることが極めて重要である。さらに責任の所在の問題もでてくるので、構想段階から法律の専門家なども含めて議論を行うことが将来的なリスクを低減することにつながるだろう。≫

今後、AI はますます進展し、それにともない AI に関連する企業・組織の社会的責任は大きくなり、「責任ある AI」が重要になります。アクセンチュアによる複数の調査によると、「責任ある AI」にしっかり対応できている企業とそうでない企業とを比較した場合、投資収益率で実に 3 倍もの差がある、といいます。
「責任ある AI」についての対策を講じている企業では自信を持って AI ビジネスに取り組み拡張・拡大が図れるのに対して、そうでない企業の場合、恐る恐る現場任せとなってしまうためです。
つまり、「責任ある AI」への取り組みはコストではなく、収益を生み出すための活動であるという認識が重要になります。

（＊）保科学世・鈴木博和『責任ある AI―「AI 倫理」戦略ハンドブック』、東洋経済新報社、2021 年刊