AI の認証制度をどう設計するか？

羽深宏樹

はじめに

自動車、医療機器、金融、そして生成 AI と、社会のあらゆる場面で AI の実装が進んだことに伴い、AI に対する認証制度の整備を求める声が高まっている。「『AI 制度に関する考え方』について」[1]等の政府文書では、「第三者認証」といった制度が度々言及されている。また、米国の”Voluntary AI Commitments”[2]においても、第三者による脆弱性の発見・報告やレッド・チーミングが高度な AI の開発者に対して求められており、こうした第三者による認証や監査を規制手法に組み込む選択肢も想定される。

認証（Certification）とは、独立した機関が、対象となる製品、プロセス、サービス、またはシステムが特定の要件を満たしていることを証明する文書（証明書）を提供することをいう[3]。AI システムは、非常に複雑なアルゴリズムからできており、外部からこれを信頼できるかどうかを判断することは難しい。また、仮にアルゴリズム自体を開示されたとしても、その機能やリスクを読み解くのは一般的には困難だ。そこで、専門性のある独立の第三者が、当該 AI システムに対して、特定の要件を満たすものであることに対する「お墨付き」を与えるニーズがあるのだ。

しかし、AI について認証制度を作ることは、容易なことではない。本稿では、AI に関する認証制度について、その課題と展望を説明する。

1．プロダクト認証とマネジメント認証

AI に限らず、ある製品やサービスについて認証制度を作る場合には、①製品やサービス自体が備えるべき性能や構造に対する認証（プロダクト認証）と、②AI システムを開発・運用する事業者の組織に対する認証（マネジメント認証）という 2 つのアプローチがある。

①のプロダクト認証は、いわゆる製品安全を扱うものだ。たとえば、日本では、疾病の診断・治療等を目的としたソフトウェア（医療機器プログラム）の製造販売について、そのリスクに応じて厚生労働大臣による承認または登録認証機関による認証を必要と

[1] AI 戦略チーム「『AI 制度に関する考え方』について」（2024 年 5 月）
https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo2-1.pdf

[2] “Voluntary AI Commitments to the White House”, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

[3] ISO ウェブサイト：https://www.iso.org/conformity-assessment.html

している[4]。また、自動運転車に搭載される自動運行装置についても、国土交通大臣によって型式指定が行われる[5]。

EU では、AI 法を、製品安全に関する一連の政策パッケージである NFL（New Legislative Framework）の一環に位置付けている。NLF には、AI 法のほかに、機械やバッテリー、医療機器規則等の製品の安全規制が含まれている[6]。AI 法では、ハイリスク AI システムに関する要件が 8 条から 15 条にかけて定められており、そこでは、リスクマネジメントシステム、データガバナンス、技術文書、記録保持、透明性、人間による監視、正確性とセキュリティなどが挙げられている。

他方、②のマネジメント認証とは、AI システムそれ自体の要件ではなく、AI システムを開発したり運用したりする事業者に課せられる義務だ。たとえば、薬機法では、医療機器等の製造業者に対して、製造管理及び品質管理の基準の遵守[7]や、製造及び試験に関する記録の保管[8]が義務付けられている。EU の AI 法では、ハイリスク AI のプロバイダに対して、品質マネジメントシステムの確立が義務付けられるほか、技術文書や品質マネジメントシステム関連文書、自動生成ログの保管などが義務付けられている[9]。

法的義務ではないが、AI マネジメントシステムについては、2023 年に、ISO/IEC 42001[10]という国際規格が発行された。これは、世界的に多くの企業が認証を取得している ISO/IEC 27001（情報セキュリティマネジメントシステム）[11]と同じ系統に属する国際規格である。

2. 現行の認証システムの課題

しかし、これらの 2 つの認証アプローチのいずれも、これを AI システムに適用するためには困難な課題が伴う。①のプロダクト認証については、AI のアルゴリズムが継続的に学習を重ね、アップデートされることから、流通開始前に認証を行うだけでは不十分であり、継続的に認証を行い続けなければならないという問題がある。また、AI システムの出力は確率的に行われるため、同じ入力を行っても、その都度出力が異なる場合もある。そのようなアルゴリズムに対して、何をもって「安全である」と評価できるのかが課題となる。公平性やプライバシーといった数値化困難な価値について、アルゴリズムがどのような水準を満たせば問題なしといえるのかの基準設定も難しい。さらには、AI システムには他の様々なシステムが接続されており、バリューチェーン上に多くの主体がいることから、AI システム自体がどこまでの信頼性を満たせばよいの

---

[4] 薬機法 23 条の 2 の 5、同 23 条の 2 の 23

[5] 道路運送車両法 41 条 1 項 20 号、75 条の 3 第 1 項

[6] https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en

[7] 薬機法施行規則第 114 条の 58、QMS 省令 83 条

[8] 薬機法施行規則第 114 条の 51、医療機器及び体外診断用医薬品の製造管理及び品質管理の基準に関する省令（平成 16 年厚生労働省令第 169 号）（以下「QMS 省令」）9 条、68 条。

[9] EU AI 法 17 条～19 条

[10] https://www.iso.org/standard/81230.html

[11] https://www.iso.org/standard/27001

かの判断も困難だ。たとえば、2023 年に施行されたニューヨーク市の採用アルゴリズムに関する規制では、年に一度のアルゴリズム監査が義務付けられているが、法律の曖昧な定義や、監査人やツール開発者、使用企業の役割に関する不明確な責任分担などによって、実効的な監査には至っていないとの指摘がある[12]。

他方、②のマネジメント認証については、AI システムを評価することなくマネジメント体制を評価するだけで、本当に AI システムの信頼を担保できるのかという点が疑問となる。たとえば、米国におけるソフトウェア医療機器の承認プロセスについて、個々の医療機器（システムレベル）ではなくそれを製造する組織（マネジメント）の要件を重視し、一定の基準を満たす企業には製品の個別承認プロセスを簡素化または省略できるようにする試みが行われた（Pre-certification Program）[13]。しかし結果的には、ソフトウェアの開発プロセスの多様性や、組織評価の基準の欠如、臨床性能やサイバーセキュリティなどデバイス固有の要素を認定する必要性などを理由として、このアプローチは有効に機能しなかったことが報告されている。

このようなプロダクト認証及びマネジメント認証に加えて、ユーザ側にも一定の専門性を求めるアプローチもある。たとえば、プログラム医療機器を使用できるのは医師に限定されるし、AI 法務サービス（リーガルテック）を利用して第三者に法的サービスを提供できるのは、原則として弁護士に限られる[14]。逆に言えば、このような専門家による使用を前提にする場合には、AI システム自体に予測不可能性などの限界があっても、社会的なリスクは限定的になる（したがって AI の積極的な活用を認めてよい）という判断もあり得るだろう。

3. プロダクト認証とマネジメント認証を統合するアプローチ

このように、AI に関するリスクは、AI システムのプロダクト認証のみで評価できるわけではなく、マネジメントシステムやユーザーのリテラシーなどを総合的に考慮して評価すべきである。そうだとすれば、認証制度もこれらを総合的に考慮した仕組みとすることが望ましい。このような問題意識から、プロダクト認証とマネジメント認証を組み合わせたジョイント認証（Joint Certification）と呼ばれる制度の検討が進んでいる[15]。

---

[12] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. “Auditing Work: Exploring the New York City algorithmic bias audit regime”. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 1107–1120. https://doi.org/10.1145/3630106.3658959

[13] U.S. FDA, “The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings” (2022), https://www.fda.gov/media/161815/download?attachment

[14] 法務省大臣官房司法法制部「ＡＩ等を用いた契約書等関連業務支援サービスの提供と弁護士法第 72 条との関係について」（2023 年 8 月）、https://www.moj.go.jp/content/001400675.pdf

[15] ISO/IEC NP 25336 Information technology — Artificial intelligence — High-level framework and guidance for the development of conformity assessment schemes for AI systems, https://standardsdevelopment.bsigroup.com/projects/9024-10625#/section
あわせて、Kimberly Lucy, ”Creating Trust in AI Through Standards: A Management System Approach” (https://jtc1info.org/wp-content/uploads/2022/06/01_05_Kimberly_ISO_IEC_AI_Workshop-Trust-in-AI-through-MSS_Kim-Lucy.pdf)も参照。

その背景には、従来のマネジメント要件が組織的な要件のみに着目しており、AI システム自体の安全性やテスト手法に対する評価が不十分であるという事情がある。たとえば、AI マネジメントシステムに関する ISO/IEC 42001 では、コンテキスト、リーダーシップ、計画、支援、運用、評価、改善という項目建てがなされているが、その構成及び各項目の内容は、一般的な品質マネジメントシステム（ISO/IEC 9001）や情報セキュリティマネジメントシステム（ISO/IEC 27001）の規格と大差あるものではない。他方で、AI ガバナンスに関する具体的な行動要件（リスクマネジメント、データガバナンス、試験・検証など）は、別紙 A の「コントロール目標及びコントロール」に規定されているが、こちらはあくまでも参照ドキュメントであり、そこで挙げられた行動を全て取ることが求められるわけではないし[16]、それらの行動をとっているかどうか自体が認証の対象となるわけではない。

それでは、別紙Aの内容も含めて認証の対象とすればよいかというと、そのような単純な話でもない。仕組みも用途も千差万別の AI システムについて、それぞれに最適なリスクマネジメント、データガバナンス、試験・検証などの手法や必要なコストは異なる。その中で、一定の機能やサービスを区切って一律の基準を決めることは可能だろうか。あるいは、各事業者に一律の基準を適用するのではなく、各事業者が自身の AI プロダクトやサービスに最適な行動要件を選択し、実践する能力を保有することを認証するというアプローチもあり得るが、その場合、何を基準にその能力を判断すればよいのだろうか。更には、認証にあたって分野横断的な専門性が必要となる中で、認証機関に必要な人材は確保できるだろうか。このように、検討しなければいけない課題は山積している。

おわりに

信頼できる AI を第三者がどのように認証することができるのかは、AIガバナンスの実装において最も重要かつ難しい課題のひとつだ。本稿でみたように、従来の認証の世界にはプロダクト認証とマネジメント認証の2つの系譜があるが、出力の予測が不可能かつアルゴリズムが頻繁に更新され、バリューチェーン上の主体も多いという難しい特徴をもつ AI プロダクトについて、いずれの手法も現状のままでは不十分である。そこで、これらの手法を統合する検討も始まっているが、具体的な統合の方法や認証機関の専門性確保などに関する課題は多い。

CFIEC の「AI の活用における課題と施策に関する研究会」では、このような課題を検討するため、医療機器や自動運転者に関する世界の既存の制度を比較しながら、あるべき認証制度の姿について国際比較を行っている。本稿はその背景について簡潔に説明したものであるが、今後必要となる分野横断的な認証制度の仕組みに関する今後の議論を少しでも喚起するものとなれば幸いである。

[16] ISO/IEC 42001 Annex A A.1