2025 年度 AI の活用における課題と施策に関する研究会報告書：

人間の監視を受けないハイリスク AI（UHAI）の評価、認証及び法制度の実現に関する論点整理

2026 年 3 月

一般財団法人 国際経済連携推進センター（CFIEC）

© 20265 CFIEC. All rights reserved.

CFIEC AI 研究会報告書：人間の監視を受けないハイリスク AI（UHAI）の評価、認証及び法制度の実現に関する論点整理

目次

エグゼクティブ・サマリー

１．人間を超える安全をもたらす“UHAI”の可能性

人工知能（AI）の能力は、私たちの想定をはるかに超えるスピードで進化している。2023 年以降、最先端の AI は医師国家試験や司法試験（短答式）で合格基準を大幅に上回る正答率を記録し、2026 年 1 月の大学入学共通テストでは総合得点率 96.9%を達成した。

こうした能力の向上は、試験や対話の場にとどまらない。医療画像診断においては、AI は複数の放射線科医と比較して肺がん・乳がんの見落とし率を大幅に低減することが実証されている。自動車分野では、自動運転システムが人間ドライバーと比べて重傷事故を 90%削減するという驚異的な成果を示している。

こうした現実を前に、「AI を人間が監視・チェックすれば安全」という従来の前提は、もはや自明ではなくなっている。実際に、AI が人間より優れた領域では、人間が AI の判断に関与することでかえってパフォーマンスが低下することを示す実験結果も出ている。優秀な AI に人間が「口出し」することで、結果が悪くなるのである。

本報告書では、「人間の直接的な監視なしに、ハイリスクな活動を自律的に行う AI」―UHAI（Unsupervised and High-risk Activities AI）―という概念を定義し、その社会実装に必要な制度の姿を描き出す。レベル 4 以上の自動運転車、医師なしで診断する AI 医療機器等がその典型例である。UHAI の社会実装は、単なる効率化にとどまらず、人間単独では到達できなかった水準の安全性を社会にもたらす可能性を持つ。しかし、そのためには、UHAI に対する「評価」「認証」「法制度」の三つの壁を越えなければならない。

２．UHAI の実装を阻む三つの制度的障壁

UHAI には、①判断過程が開発者にも説明できない「ブラックボックス性」、②学習により挙動が変化し続ける「動的変化」、③安全の閾値をどこに設定するかが未定義の「ゴール設定の困難」という三つの技術的特性がある。これが、従来の制度について以下の課題をもたらす。

① 評価基準の欠如： AI の「ブラックボックス性」のため、従来の製品検査の手法では安全性を証明できない。

② 認証制度の不備： 評価基準ができたとしても、独立した第三者が AI の安全性を「認証」するための仕組みや体制が整っていない。

③ 法制度の構造的限界： 現行の規制・法執行・責任法制はいずれも「最終的には人間が判断・介入できる」ことを前提に設計されている。

３．現行制度の到達点と課題（ケーススタディ）

本報告書は、医療機器 AI（日米の IDATEN・PCCP 制度等）と自動運転（日本の特定自動運行制度、英国の自動運転車法、カリフォルニア州の配備許可制度）のケーススタディから、三つの共通した方向性を析出した。第一に、AI の「中身」の解明を断念し、「振る舞い」の検証と「組織ガバナンス」の評価を組み合わせる二段階評価。第二に、一時点の審査から製品ライフサイクル全体を通じた継続的管理への移行。第三に、適切な管理体制の維持をセーフハーバー（免責の根拠）とする法的枠組みの構築である。ただし、いずれの精度にも未解決の課題があり、UHAI を巡る評価・認証・法制度の整備は発展途上である。

４．本報告書の提言

以上の課題を踏まえ、本報告書は以下の三つの方向性を提言する。

（１）UHAI の評価について

「ブラックボックス性」「動的変化」「ゴール設定の困難性」という UHAI の三つの技術的特性に対応する新たな評価パラダイムを構築する。ブラックボックス性に対しては、AI の内部を解明するのではなく、形式手法やレッドチーミング等の新たな V&V 手法を活用して AI の「振る舞い」を外側から検証する技術を標準化する。動的変化に対しては、出荷前の検証と市販後のリアルワールドパフォーマンス監視を一体的に設計し、ライフサイクル全体を通じた継続的評価の仕組みを整備する。ゴール設定の困難に対しては、「絶対的な安全」を求めるのではなく、従来の人間によるオペレーションとの相対的かつ統計的な比較を軸に、多様なステークホルダーが参加する透明なプロセスでリスク許容水準を策定し、その合意を評価基準の出発点とする。

（２）UHAI に対する認証について

上記の評価基準を実効的に担保するための認証制度を構築すべきである。ブラックボックス性に対しては、製品の入出力検証だけでなく、それを開発・運用する組織のガバナンス体制を併せて審査する「ジョイント認証」の枠組みを構築する。動的変化に対しては、変更のたびに一から審査をやり直すのではなく、変更管理のプロセスそのものを事前に承認するアプローチを発展させる。ゴール設定の困難に対しては、社会的に合意された基準を認証の判定基準に接続し、認証機関がそれを独立に検証しうる能力を構築する。

（３）UHAI に関する法制度について

評価と認証の成果を法的に受容し、事業者の予見可能性を確保することが重要である。ブラックボックス性に対しては、内部構造の説明に代えて安全動作範囲の

技術的評価を許認可の要件として受容し、適切な組織体制の維持をセーフハーバーとして法的に位置づける。動的変化に対しては、事業者に継続的な安全管理と自主報告を義務付けつつ、情報提供への適切なインセンティブ付けを行うことによってその実効性を確保する。ゴール設定の困難に対しては、社会的に合意された基準を達成していることを事業者が証明できるようなテスト環境や認証制度を整備する。

これら三層は個別に設計するのではなく、統合的に設計・運用される必要がある。すなわち、評価基準の合意が認証の基盤となり、認証の取得がセーフハーバーの根拠となり、セーフハーバーの明確化が事業者の予見可能性を高めてイノベーションを促進する。

UHAI は、人間の能力の限界を超え、社会により大きな安全と便益をもたらす可能性を秘めている。その可能性を現実にするために、私たちは「人間が監視すれば安全」という従来の前提を超え、技術の進化に適応しうる新たな制度の枠組みを構想しなければならない。本報告書で整理した論点が、その議論の出発点となることを期待する。

＜表：UHAI の特徴（横軸）及びそれに対応する制度設計の方向性（縦軸）＞

	ブラックボックス性	動的変化	ゴール設定の困難性
評価	AI の内部解明を断念し、形式手法・レッドチーミング等の新たな V&V 手法で AI の「**振る舞い」を外側から検証**する技術を標準化	出荷前検証と市販後の**リアルワールドパフォーマンス（RWP）監視を一体設計**し、ライフサイクル全体を通じた継続的評価を整備	「絶対的な安全」ではなく、人間によるオペレーションとの**相対的・統計的比較**を軸に、多様なステークホルダー参加の透明なプロセスでリスク許容水準を合意
認証	製品の入出力検証＋組織のガバナンス体制を**セットで審査する「ジョイント認証」**の枠組みを構築	変更のたびに再審査するのではなく、**変更管理プロセスそのものを事前承認**するアプローチを発展	社会的に合意された基準を**認証判定基準に接続**し、認証機関がそれを独立に検証しうる能力を構築
法制度	内部構造の説明に代え、**安全動作範囲の技術的保証を許認可要件**として受容。適切な組織体制の維持を**セーフハーバー**として法的に位置づけ	事業者に**継続的安全管理と自主報告を義務**付けつつ、情報提供への適切なインセンティブ付けにより実効性を確保	合意された基準の達成を事業者が証明できる**テスト環境・認証制度を整備**し、法的予見可能性を確保するセーフハーバーを構築

1. 本報告書の背景と目的 ― 人間を超える安全をもたらす“UHAI”の可能性

1.1 人工知能の脅威的な進歩

我々は今、テクノロジーの歴史における特異な瞬間を生きている。

1950 年、コンピュータ科学の父アラン・チューリングは、「人間の審判がテキスト対話を通じて人間と機械を区別できなければ、その機械には知性がある、とみなす」という有名な「チューリングテスト」を提唱した。以来 70 余年、このテストは AI 研究の究極の目標であり、同時に「超えられない壁」の象徴でもあった。しかし 2023 年以降の研究において、GPT-4 が審判の過半数を人間と誤認させることに成功し、さらに 2025 年には GPT-4.5 が 73%の割合で人間と判定されるに至った [1]。

高度な専門知を備えた人間を選別するための試験においても、AI は合格点を大幅に上回る成績を記録し始めている。たとえば 2026 年 1 月の大学入学共通テストでは、最新世代の AI が主要科目で満点を連発し、総合得点率 96.9%という驚異的なスコアを叩き出した [2]。医師国家試験においても、AI は難解な症例判断を含む問題で現役医師に匹敵する正答率を維持し、安定して合格圏内に位置している [3]。さらに 2025 年の司法試験（短答式）では、人間のトップ層に並ぶ正答率 96%以上を記録した [4]。

AI が人間以上のパフォーマンスを発揮できる領域は、対話や試験問題の枠にとどまらない。我々の生命や生活に直結し、そのために従来規制で厳格に規律されて

---

[1] Jones, C. & Bergen, B. (2024). People cannot distinguish GPT-4 from a human in a Turing test. arXiv:2405.08007. 同研究では GPT-4 が審判の 54%に対して人間と誤認させることに成功した。さらに、Jones, C. & Bergen, B. (2025). Large Language Models Pass the Turing Test. arXiv:2503.23674 では、GPT-4.5 が 73%の割合で人間と判定された。

[2] LifePrompt 社・日本経済新聞社共同検証（2026 年 1 月 20 日公表）。OpenAI の GPT-5.2 Thinking が 2026 年度大学入学共通テスト 15 科目で得点率 96.9%、9 科目で満点を達成。日本経済新聞 2026 年 1 月 21 日付「大学共通テスト オープン AI、9 科目満点 得点率 97%」参照。

[3] Takagi, S., et al. (2023). Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Medical Education, 9, e48002; Nomura, A., et al. (2024). Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan. PLOS Digital Health, 3(1), e0000433. 第 119 回医師国家試験（2025 年）では、OpenAI o1 が必修問題で得点率 98%を達成した（メディックメディア検証）。

[4] 弁護士ドットコム株式会社「法律特化 AI『Legal Brain』が 2025 年度司法試験（短答式）で正答率 96.5%を記録」（2025 年 11 月 12 日）。175 点満点中 169 点を獲得し、受験者最高得点（167 点）を上回った。

きた「高リスクな判断」においてさえ、AI は人間より正確で信頼に足る存在となりつつある。以下では、医療と自動運転における具体的な事例を紹介する。

1.1.1 医療画像診断

医療画像診断は、AI が人間の専門家を上回ることが最も早く実証された領域の一つである。

肺がん検診では、2019 年に Google AI と Northwestern 大学の共同研究チームが、低線量 CT スキャンにおいて AI が経験豊富な放射線科医 6 名全員を上回る診断精度を達成したことを報告した。AI は、がんの見落とし（偽陰性）を 5%、誤検出（偽陽性）を 11%削減した [5]。

乳がん検診でも同様の結果が出ている。2020 年、Google Health、DeepMind、英国 NHS の共同チームは、マンモグラフィにおいて AI が偽陽性率を最大 5.7%、偽陰性率を最大 9.4%削減したことを Nature 誌に発表した。6 名の放射線科医との直接比較では、AI が全員を上回った [6]。

米国では、完全自律型の AI 診断が規制当局に承認され始めている。2018 年、FDA は糖尿病網膜症診断 AI「IDx-DR」を、医師の判断を介さずに AI が単独で診断結果を出力することを認める形で承認した。これは、「特定領域において AI が人間の専門家なしに医療判断を行う」ことを規制当局が公式に認めた最初の事例である [7]。

1.1.2 自動運転：人間ドライバーより安全な AI

自動運転の分野でも、昨今では機械による運転の安全性が人間によるそれを大幅に上回るようになってきている。

Google 傘下の Waymo は、自社の自動運転車両の安全性に関する査読付き論文を継続的に発表している。2025 年時点で累計 1 億 2,700 万マイル（約 2 億キロメートル、地球約 5,000 周分）以上の公道走行データに基づく最新の分析によれば、Waymo の自動運転車両は人間ドライバーと比較して、重傷以上の事故を 90%、

---

[5] Ardila, D., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25, 954–961.

[6] McKinney, S.M., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89–94.

[7] Abràmoff, M.D., et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1, 39.

負傷事故全体を 81%、歩行者負傷事故を 92%削減している [8]。すなわち、自動運転システムは一定の運行条件下において、人間ドライバーよりも遥かに安全に車両を運行できるのである。

1.1.3 「人間によるモニタリング」という幻想

このように、規制の対象となるようなハイリスク分野において、AI が平均的な有資格者（人間）より優れたスコアを記録する場合、ルールはどのようにデザインされるべきだろうか。

最も単純な解決策は、従来の規制を維持したまま、資格をもった人間が AI を補助的なツールとして用いることを認めることである。すなわち、人間が AI の挙動を監視するモデルである。こうしておけば、通常は優れた判断を行う AI の能力を存分に発揮させつつ、万が一 AI が不適切な挙動をしたときには、人間が介入して事故を防ぐことができる。

一見すると有効に見えるこの方法には、実は大きな問題がある。それは、人間と AI が協働する場合に、かえってパフォーマンスが低下する場合があるということである。

人間と AI の協働に関する 106 の実験研究（被験者総数 16,000 人以上）を分析した実験では、人間-AI 協働は、人間単独・AI 単独の両方を上回る「シナジー効果」を示さなかった。特に、AI が人間より優れた能力を持つ場合、人間の参加がパフォーマンスを大幅に低下させた [9]。

このメカニズムの一つは「Automation Bias（自動化バイアス）」である。2023 年の研究では、AI が誤った診断を示した場合、経験の浅い医師の正答率は 19.8%にまで急落し、経験豊富な医師でも 45.5%に低下することが示された [10]。AI が間違えると、医師もつられて間違える。しかも、この傾向は訓練や注意喚起では防止できないことが、複数の研究で示されている。

これらの事実は、「人間がダブルチェックすれば安全」という直感的な想定が、少なくとも特定の条件下では成り立たないことを示している。優秀な AI に人間が「口出し」することで、かえって結果が悪くなるのである。

---

[8] Kusano, K.D., et al. (2025). Comparison of Waymo Rider-Only Crash Rates by Crash Type to Human Benchmarks at 56.7 Million Miles. *Traffic Injury Prevention*, 26(sup1), S8–S20.

[9] Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8, 2293-2303.

[10] Dratsch, T. et al. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4).

また、自動運転の文脈では、そもそも咄嗟の危険判断に対して人間が対応する余裕がない場面も多い。

このような条件のもとで、なお従来の規制アプローチを維持し、人間に AI のモニタリングを義務付けることは、かえって安全性を低下させるばかりではなく、監視役の人間に対して実効性の伴わない過大な責任を負わせ、結果として社会全体の便益を損なう「形骸化した安全」に固執することにつながりかねない。

1.2 「人間の直接的な監視なしにハイリスクな活動を自律的に行う AI」(UHAI)に関する制度整備の必要性

1.2.1 UHAI とは何か

以上の問題意識に基づき、本報告書では、「人間の直接的な監視なしに、ハイリスクな活動を自律的に行う AI」（UHAI: Unsupervised and high-risk activities AI）という概念を用いて、これに関する認証制度及び法制度の在り方を検討する。まずは、UHAI という概念について整理する。

UHAI の第一の特徴は、「ハイリスクな活動」を行う AI ということである。ハイリスクな活動とは、その AI の判断と実行が、市民の生命、身体、財産、あるいは基本的人権といった法的保護利益に対して、直接的かつ重大な影響を及ぼしうる活動を意味する。具体的には、公道におけるレベル 4 以上の自動運転車、診断や治療計画を自律的に実行する医療機器、金融市場の安定を左右するアルゴリズム取引システム、そして重要インフラの自動制御 AI などがこれに該当する。何がハイリスクに該当するかの考え方は、国や文化によって異なるが（たとえば、EU の AI 法では、既存の規制に加えて、AI による遠隔生体識別やスコアリングなどをハイリスク領域と指定している [11]）、本稿ではその点には立ち入らない。一般的に、安全規制の対象となっている判断については、社会的に「ハイリスク」とみなされると理解してよい。

UHAI の第二の特徴は、このようなハイリスクな決定や活動を、直接的な人間の監視なしに AI が自律的に行うということである。たとえハイリスクな挙動を行う AI であっても、人間の補助ツールとして使われるにすぎないもの—例えば、AI が提示した診断候補を医師が最終的に判断する診断支援システムや、人間のドライバーが常に周囲を監視し緊急時に介入するレベル 2 の自動運転技術など—は UHAI に含まれない。

[11] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (AI Act), Annex III (High-Risk AI Systems Referred to in Article 6(2)).

要するに、UHAI は、ハイリスクなシステムの運用における「オペレーター」の役割を、人間から AI システムへと移行させるものである。この移行は、単なる業務効率の向上にとどまらず、従来人間によるオペレーションだけでは到達できなかった高水準の安全性を実現することにつながる。すなわち、UHAI の社会実装を一定の場合に法的に認めることによって、人間の認知や能力の限界を超えて AI がもつ潜在的可能性を引き出し、社会により大きな安全性や便益をもたらすことが想定されるのである。

1.2.2 UHAI 実装に立ちはだかる「評価」「認証」「法制度」の壁

しかし、UHAI を社会実装するためには、越えなければいけない壁がある。

第一に、UHAI がハイリスク領域を扱うものである以上、どのような要件を満たす AI であれば信頼できるかについての「評価基準」がなければならない。しかし、AI が持つ高度な自律性、そして複雑な機械学習がもたらす予測不可能性や説明不可性（ブラックボックス性）といった複合的な特性は、消費者、規制当局、さらには開発者自身でさえ、その安全性と信頼性を完全に評価することが極めて困難にしている。アルゴリズム自体を評価することで安全性を確認することが事実上困難である場合に、何を基準に信頼を判断すればよいのかが技術的な問題となる。

第二に、ハイリスク分野を扱う UHAI については、実際にシステムが上記の基準を満たすかどうかについて、単に開発者自身が検査するだけでは足りず、専門的な第三者が評価して「認証」が必要となる場合も多い。認証とは、独立した機関が、対象となる製品、プロセス、サービス、またはシステム等が特定の要件を満たしていることを評価し（適合性評価）、これを証明する文書（証明書）を提供することをいう。UHAI のシステムとしての複雑性を考えると、開発者でない第三者が、そのような評価や保証を行うことは極めて困難である。そのため、UHAI の認証制度を確立する場合、誰が、何を対象として、いつ、どのように認証を付与することができるかも問題となる。

第三に、UHAI を、「法制度」との関係でどのように扱うかという問題もある。現行の規制は原則として、AI システムを使う場合であっても、あくまで人間の判断をサポートする範囲で AI システムを用いることを想定している。すなわち、最終的な責任者はあくまで人間であることを前提としている。たとえば、EU AI 法では、ハイリスク AI システムについて人間による監視（Human Oversight）を義務付けており [12]、UHAI の運用は原則として許容されていない。また、日本においても、医療分野では、厚生労働省の行政通知である「医政医発 1219 第 1 号（平成

[12] Regulation (EU) 2024/1689 (AI Act), Article 14 (Human Oversight).

30 年 12 月 19 日）」において、[13]AI を用いた診療支援が行われる場合であっても「診断、治療等の行為の主体は医師であり、医師が最終的な判断の責任を負う」という原則が明確に示されている。自動運転車についても、2022 年の道路交通法改正 [14] により法律上はレベル 4 相当の「特定自動運行」が可能とされているが、その許可には遠隔監視を行う「特定自動運行主任者」の配置が義務付けられており、完全に人間の介在なしに AI が自律的に運行する形態は想定されていない。このように、現行の法制度下では、AI の自律的な判断のみに依拠する UHAI の実装には、依然として大きな制度的障壁が存在している。

1.3 本報告書が前提とする UHAI の技術的特徴

本報告書は、上記のような課題を踏まえて、UHAI を社会実装するための信頼基準の在り方を検討すると共に、UHAI に関する認証制度及び法制度のあるべき姿を描き出すことを目的とする。

その際、本稿では、UHAI の技術的な特性として以下の 3 点に着目する。いずれも、UHAI の「評価」、「認証」、および「法制度」のいずれにも波及する、UHAI に固有の根本的な技術的特性である。

1.3.1 ブラックボックス性

深層学習を用いた AI は、多数のパラメータ間の非線形な相互作用によって判断を導出するため、その判断プロセスが開発者を含む人間には解釈不能な「ブラックボックス」となる。これは、従来の工業製品のように仕様書に基づいて内部動作の適正性を論理的に検証するアプローチが、AI に対しては原理的に適用困難であることを意味する。

1.3.2. 動的変化

AI モデルは、継続的な学習、データ入力の変動、および外部環境の変化によって、運用中にその挙動が変化しうる。この動的変化は、ソフトウェアの計画的なアップデートとは異なり、連続的かつ非線形に生起する場合があるため、ある時点での安全性評価（スナップショット評価）が将来にわたって有効であることを保証できない。さらに、UHAI においては人間の監督者が不在であるため、動的変化に伴う性能劣化や予期せぬ挙動の変容が長期間にわたり検知されないリスクがある。

---

[13] 厚生労働省医政局医事課長通知「人工知能（AI）を用いた診断、治療等の支援を行うプログラムの利用と医師法第 17 条の規定との関係について」（医政医発 1219 第 1 号、平成 30 年 12 月 19 日）。

[14] 道路交通法の一部を改正する法律（令和 4 年法律第 32 号）。特定自動運行に関する規定は第 75 条の 12 以下。

1.3.3 ゴール設定の困難性

UHAI は、確率と統計をベースとするシステムである以上、「完全な安全性」を保証することは原理的に不可能であり、一定程度のリスクを内在することを避けることはできない。そのため、「どの水準の安全性を達成すべきか」を社会として定義する必要があるが、その水準については絶対的な正解があるわけではなく、文化的背景や社会的ニーズによって受容可能性は多様である。さらに、物理的な安全性に加えて、「公平性（差別的挙動の防止）」や「説明可能性（判断根拠の提示）」といった定性的要件をどのように測定・評価すべきかという方法論自体が未確立であるため、評価の「ものさし」を設定すること自体が困難である。

1.4 本報告書の構成

本報告書では、上記のような UHAI の技術的な特徴を踏まえ、以下のように分析を進める。まず、第 2 章では、UHAI の性質と現行制度のギャップについて、「評価」「認証」「法制度」の 3 つの観点から分析する。第 3 章では、これらのギャップを解消する手がかりを掴むために、医療機器 AI（SaMD）および自動運転（AV）の二領域を対象に、主要国の規制モデルが UHAI の課題にどのように対応しているかをケーススタディとして検討する。第 4 章では、これらの分析を踏まえ、評価基準の構築、新たな認証制度の導入、および法制度の再設計について、今後の制度構築に向けた論点整理を行う。

2. UHAI の実装を阻む三つの制度的障壁

以下では、UHAI を社会に実装する際に存在する、UHAI の性質と現行制度のギャップについて、UHAI に関する「評価」(2.1)、「認証」(2.2)「法制度」(2.3)という 3 つの観点から分析する。

2.1. UHAI の評価に関する課題

UHAI の認証を困難にしている根本的な要因は、何を基準に「合格」とするかという評価基準（ものさし）の設定自体が困難な点にある。

2.1.1 AI プロダクト評価の難しさ

従来の工業製品（電気機器等）の認証は、事前に定められた仕様や性能基準に製品が適合しているかを検査する「仕様適合性」のアプローチが一般的であった。しかし、1.3 でみたような「ブラックボックス性」「動的変化」「ゴール設定の困難性」という特徴をもつ UHAI に対してこのアプローチを適用することには以下の限界がある。

第一に、ブラックボックス性については、UHAI の判断プロセスが人間には解釈不能であるため、仕様書通りに動作しているかを論理的に検証することが困難である。結果として、網羅的なテストが不可能であり、製品の安全性を完全には保証できない。

第二に、動的変化については、ある時点での「スナップショット」としての評価結果が、運用中の UHAI の安全性を将来にわたって保証するものではないという問題をもたらす。

第三に、ゴール設定の困難性については、まさに UHAI の信頼性を評価するための評価指標や評価手法自体の策定が困難であることを意味する。

2.1.2 オペレーター（組織）に対するの評価の限界

このようなプロダクト評価の限界を補完するものとして、プロダクト自体ではなくこれを開発・運用する組織に対して評価を行うアプローチがある。ISO/IEC 42001 に基づくマネジメントシステム認証（組織認証）がこれに該当する。これは、組織が AI のリスクを特定・分析し、継続的に改善する PDCA サイクル等の「組織のガバナンス体制」を評価するものであり、AI の信頼性を確保するための基盤として極めて重要な役割を果たす。

しかし、UHAI の認証においては、この組織認証のアプローチ「単独」では以下の限界が残る。 マネジメントシステム認証は、あくまで組織が「適切なプロセスで開発・運用しているか」を保証するものであり、そのプロセスから生み出された「個別の UHAI 製品が、特定の状況下で安全に挙動すること」までを直接的に保証するものではない。特に、人間の介入なしに自律的に判断を行う UHAI においては、組織がどれほど健全であっても、ブラックボックスな製品自体が予期せぬ挙動（ハルシネーションや事故等）を引き起こすリスクを完全に排除することはできない。 したがって、UHAI の信頼性を担保するためには、組織の管理能力を評価するだけでなく、その管理下にある製品固有のリスク対策が有効に機能しているかを確かめる視点が不可欠となる。

このように、プロダクト認証およびマネジメントシステム認証のいずれの単独アプローチも、UHAI が内包する「ブラックボックス性」「動的変化」「ゴール・評価手法設定の困難」といったリスクの特徴に対応するには不十分であるといえる。

2.2 UHAI への認証付与に関する課題

仮に UHAI の評価基準が策定できたとしても、それを誰が、いつ、どのように認証するのかという認証制度の構築には、なお高いハードルが存在する。

第一に、情報の非対称性と第三者評価の限界である。AI システムのリスクや挙動を最も理解しているのは開発者（提供者）自身であり、外部の認証機関が限られた情報と時間の中で、ブラックボックスな AI の中身を正確に審査することは極めて困難である。 第二に、認証のタイミング（Time-to-Market）**の問題**である。AI 技術の進化やモデルの更新頻度は極めて速い。認証に数ヶ月を要するような従来の制度設計では、認証を取得した頃にはモデルが更新されているか、技術的に陳腐化している恐れがある。 第三に、**認証対象の切り出しの困難**である。UHAI は単独で存在するのではなく、ハードウェア（例：自動車・医療機器）やクラウド環境、API 連携する外部サービスと複合的に動作する。この「System of Systems」において、どこまでを認証の範囲とするかの線引きは技術的・法的に未解決な課題である。

2.3 UHAI の法制度に関する課題

仮に何らかの評価基準が確立され、認証制度が整備されたとしても、なお解決されない問題がある。事前の規制で UHAI をどこまで統制できるのか。規制の枠組みが設けられたとして、それを実効的に執行できるのか。そして UHAI が実際に損害を引き起こした場合、その法的責任は誰がどのように負うのか。これらは、評価・認証の問題とは独立に存在する法制度固有の構造的課題である。

日本政府は、AI 技術の急速な発展に対応するため、硬直的な事前規制ではなく、アジャイル・ガバナンスの考え方に基づく柔軟なガバナンスモデルを志向してきた [15]。さらに近時は、AI 利活用における民事責任の解釈適用についても、現行法の枠組みの中での対応可能性が検討されている [16]。これらの政策的努力は、AI 一般の法制度整備にとって重要な基盤を提供するものである。

以下では、日本政府が示してきたそのような方向性も踏まえながら、事前規制（2.3.1）、法執行（2.3.2）、責任法制（2.3.3）の三つの次元から、この問題を検討する。

---

[15] 経済産業省 Society5.0 における新たなガバナンスモデル検討会「GOVERNANCE INNOVATION——Society5.0 の実現に向けた法とアーキテクチャのリ・デザイン」（2020 年）、同「GOVERNANCE INNOVATION Ver.2: アジャイル・ガバナンスのデザインと実装に向けて」（2021 年）、同「GOVERNANCE INNOVATION Ver.3: アジャイル・ガバナンスの概要と現状」（2022 年）。

[16] 経済産業省「AI 利活用における民事責任の在り方に関する研究会」（2025 年 8 月～）。同研究会は、AI 利活用における民事責任の解釈適用に関する手引きの策定を目指し、現行法（不法行為、製造物責任法、債務不履行等）の解釈適用の整理を進めている。

2.3.1 事前規制（上市規制）に関する課題

事前規制は、いずれの法域においても、対象の仕様が事前に確定可能であり、規制当局がその内容を審査しうることを前提とする。しかし、「ブラックボックス性」「動的変化」「ゴール・評価手法設定の困難」という UHAI の三つの特性は、この枠組みの限界を露呈させるものである。

第一に、ブラックボックス性が許認可制度の前提を揺るがす。許認可は、申請者が対象システムの仕様を明示し、その安全性を説明できることを前提とする。しかし、UHAI の判断過程は開発者自身にとっても説明が困難であり、仕様の確定という行為自体が成立しない場合がある。UHAI においては、問題は単に仕様が不明確であることではない。むしろ、運用過程において行動空間自体が生成的に拡張されうるため、事前に確定可能な仕様という概念そのものが成立しない可能性がある。人間が運用に介在する AI であれば、「最終的には人間が確認・判断する」という安全弁が機能しうるが、UHAI にはこの経路が存在しない。仕様を確定できない対象に対して許認可を付与する法的根拠は、現行制度には見当たらない。

第二に、従来の許認可は、特定の時点におけるシステムの状態に対する適合性判断であったが、UHAI の動的変化はその有効性を著しく限定する。これは単に「更新頻度を高めれば解決する」という技術的問題ではない。UHAI においては、変化が連続的かつ非線形に生起するため、許認可という制度形式が前提とする「ある時点における状態固定」という構造自体が制度的前提として成立しにくい。有人監督の AI であれば、運用者が「最近の出力に違和感がある」と気づきうるが、UHAI ではそのような監視主体が不在であるため、乖離が長期間にわたり検知されないリスクがある。「いつまで有効な認可か」という時間軸の問題に、現行の許認可制度は回答を持たない。

第三に、ゴール設定の困難が規制基準の策定を阻害する。2.1 で論じたように、UHAI の安全性の閾値自体が定義困難である。ガイドラインの実効性は、遵守すべき基準が明確であることを前提とするが、UHAI の能力は領域横断的かつ文脈依存的であり、一律の基準設定になじまない。しかも、基準の不足を人間の判断で逐次補完することもできない。

以上のとおり、仕様の不確定、時間的有効性の喪失、基準の不在という三つの問題が重なることで、事前規制の前提が構造的に成り立たなくなる。さらに、UHAI には、「人間による最後の安全弁」が存在しないがゆえに、これらの問題は一層深刻化する。

2.3.2 法執行（エンフォースメント）に関する課題

UHAI に対する法執行にも、困難な課題が存在する。

第一に、ブラックボックス性が監視の実効性を損なう。人が監督する AI であれば、運用者が異常を感知し報告するという経路を通じて、執行機関は間接的にシステムの状態を把握しうる。UHAI にはこの内部告発的な経路が存在せず、外部からの監視が唯一の手段となるが、その外部監視自体がブラックボックス性により制約される。従来の法執行における立入検査や報告徴求は、対象の内部状態が一定程度可視であることを前提とする。UHAI の内部状態は開発者にとっても不透明であり、執行機関が外部から「何が起きているか」を把握する手段は極めて限られる。外部から観察可能な入出力のモニタリングのみでは、内部の判断過程の適切性を評価することはできない。

第二に、動的変化が違反認定を困難にする。「違反」の認定には違反時点の特定が不可欠であるが、UHAI は段階的かつ連続的に変化するため、「いつから不適合になったのか」に明確な答えがない場合がある。さらに、人間の監督者が不在であるがゆえに、不適合状態が長期間にわたって蓄積されても発見が遅れる構造的リスクがある。これは、問題が顕在化した時点では既に被害が拡大しているという事態を招きかねない。

第三に、ゴール設定の困難が執行設計の基礎を欠如させる。何を監視すべきか、どの指標で判断すべきか、どの頻度で検査すべきか——法執行の具体的設計に必要なパラメータがいずれも未確定である。加えて、執行機関自身が UHAI の技術的評価を行う能力を十分に有していない現状においては、実質的な監視は著しく困難であるといわざるをえない。

このように、事前規制によって規則が設けられたとしても、それを実効的に執行する手段が追いついていないのが現状である。規制と執行の乖離は、法制度の実効性を根本から損なう。

2.3.3 責任に関する課題

UHAI が引き起こす損害に対する事後的な責任の在り方もまた、UHAI に関する法制として重要な位置を占める。そして、既に述べた事前規制とその執行の在り方と同様に、UHAI の三つの構造的特性は、責任の在り方についても様々な困難を突き付ける。

第一に、ブラックボックス性が因果関係の立証を困難にする。不法行為責任の成立には、被害者が加害行為と損害との間の因果関係を立証する必要があるが、

UHAI においてはそもそも AI の判断過程を検証することが専門家にとってすら困難である。さらに、有人監督の AI であれば、最終的には人間のミスと結果との因果関係の問題として整理することもできるが、UHAI の場合はそのような擬制を行うこともできない。因果の鎖が完全にブラックボックスの内部に閉じ込められるのである。EU では、この問題に対応するため、ハイリスク AI システムに関する因果関係を推定する法案が提案されていたが（AI Liability Directive）、同指令案は、加盟国間の合意が得られず、2025 年 10 月に正式に撤回された [17]。

第二に、動的変化が既存の欠陥概念との齟齬を生じさせる。製造物責任法は「引渡し時」の欠陥を要件としている [18]。しかし、UHAI は出荷後も学習と変化を継続するため、引渡し時には問題がなくとも、運用の過程で性能が劣化し、あるいは予期しない判断傾向を獲得しうる。「いつの時点の状態を基準に欠陥を判断するか」について、現行法は明確な回答を持たない。

第三に、ゴール設定の困難が責任範囲の画定を困難にさせる。過失の認定には「予見可能性」と「結果回避義務」の確定が必要であるが、2.2 節で論じたように UHAI の安全性の評価方法自体が未確立であるため、何を予見すべきであったか、どのような結果を回避すべきだったかの判断が困難である。さらに、AI のサプライチェーン上には開発者・提供者・利用者といった様々な主体がいるため、それぞれの主体ごとに予見可能性と結果回避義務の範囲を決めるのは極めて困難である。

2.3.4 小括

本節の検討を通じて明らかになったのは、UHAI が既存の法制度の三つの局面において構造的な課題を生じさせるということである。UHAI が提起するのは、規制の不足ではなく、人間によるリスクのコントロールという前提が崩れるという根本的な問題である。事前規制においては、仕様の不確定性、基準の不在、および時間的有効性の喪失により、規制の前提そのものが成立しない。法執行においては、システム内部の不可視性、違反時点の特定困難、および執行基準の策定の難しさにより、規制が存在したとしてもそれを実効化する手段を欠く。責任法制に

[17] European Commission, Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final (September 2022). 同指令案は、ハイリスク AI に関する因果関係の推定規定や証拠開示命令等を含んでいたが、加盟国間の合意が得られず、欧州委員会 2025 年作業計画（COM(2025) 45）の Annex IV において撤回が表明され、Official Journal C/2025/5423（2025 年 10 月 6 日）により正式に撤回された。

[18] 製造物責任法第 2 条第 2 項は「欠陥」を「当該製造物の特性、その通常予見される使用形態、その製造業者等が当該製造物を引き渡した時期その他の当該製造物に係る事情を考慮して、当該製造物が通常有すべき安全性を欠いていること」と定義する。「引き渡した時期」が基準時点とされている点が UHAI との関係で問題となる。

おいては、因果関係の立証困難、欠陥概念と AI の動的性質の齟齬、責任の範囲の画定の難しさにより、被害者救済の経路が不安定になる。

これら三つの次元の課題は独立に存在するのではなく、相互に連動している。事前規制が機能しなければ、合法・違法の境界が不明確なまま法執行の対象を画定できない。法執行が機能しなければ、違反が是正されないまま損害が生じ、その負担はすべて責任法制に転嫁される。しかし責任法制もまた、因果関係の立証困難や責任範囲のにより機能不全に陥るため、被害者救済の経路が閉ざされる。

この法制度内部の悪循環は、2.1 節・2.2 節で検討した評価・認証の課題とも連動する。評価基準が不確定であれば認証が困難となり、認証が困難であれば法的なセーフハーバーも成立しないため、責任リスクはさらに拡大する。

この悪循環を断ち切るためには、先端的 AI を巡る法制度について、国内外で現在どのような取組が行われているかを把握することが重要である。そこで、第 3 章では医療 AI 及び自動運転等に関するケーススタディを行う。

3. 現行制度の到達点と課題（ケーススタディ）

前章で指摘した通り、UHAI は、「ブラックボックス性」、「動的変化」、および「ゴール・評価手法設定の困難」という三つの本質的課題を浮き彫りにしている。

本章では、代表的なハイリスク領域である「医療機器（SaMD）」および「自動運転（AV）」の二領域を取り上げ、主要国の規制モデルがこれら三つの課題に対してどのように向き合っているかを詳細に観察・分析する。

分析の方法として、本章ではまず 3.1 および 3.2 において、各事例の制度対応を「ブラックボックス性」「動的変化」「ゴール・評価手法設定の困難」という三つの技術的特性を軸として検討する。これは、同一の技術的課題に対して各国がいかに異なるアプローチを採っているかを比較可能にするためである。その上で 3.3 において、これらの知見を第 2 章で提起した三つの制度的課題――「評価」「認証」「法制度」――に翻訳し、第 4 章における提言への架橋とする。

3.1 ＜事例 1 ＞AI を用いた医療機器の認証

本報告書冒頭の事例でも述べたとおり、医療機器ソフトウェア（Software as a Medical Device: SaMD）――すなわち、ハードウェアに組み込まれるのではなく、ソフトウェアそれ自体が診断や治療の判断を行う医療機器――の領域は、AI の活用が最も早くから規制上の論点となった分野である。2025 年時点で、FDA が認証

した AI 搭載医療機器は累計 1,300 件を超え、年間承認数は 331 件に達している[19]。

本節では、日米の制度を中心に、SaMD 規制が UHAI の三つの課題にどう対処しているかを分析するとともに、組織認証と製品認証の関係についても検討する。

3.1.1 制度の概要

日本：IDATEN 制度

日本における AI 医療機器の承認審査は、従来の医療機器規制と同じ枠組みに基づいている。すなわち、機器の内部構造やメカニズムを直接検証するのではなく、臨床性能試験（臨床データを用いた診断精度等の評価）と非臨床性能試験（ベンチテストや既知のデータセットを用いた性能評価）の組み合わせによって安全性・有効性を判断する、「入力に対してどのような出力を返すか」という振る舞いの検証が基本的なアプローチである。この枠組みは、従来の医療機器では「内部構造が安定している」という暗黙の前提のもとで十分に機能してきた。しかし、AI においては内部がブラックボックスであるだけでなく、学習により挙動が変化しうるため、同じ評価枠組みでは変化後の安全性を保証できないという構造的問題が生じる。この問題に対応するために導入されたのが、2019 年薬機法改正による IDATEN 制度である。

2019 年薬機法改正により導入された承認後変更管理実施計画書（PACMP）制度、通称 IDATEN（Improvement Design within Approval for Timely Evaluation and Notice）は、AI を含む SaMD の変更管理を迅速化するための枠組みである。開発者は初回承認時に将来のアルゴリズム変更計画を提出し、PMDA（独立行政法人医薬品医療機器総合機構）の確認を経て、計画範囲内の変更については軽微変更届出のみで実施可能とする。この制度の核心は、AI アルゴリズムの変更を「事前に計画された範囲内の変更」として扱うことで、変更のたびに最初から承認手続をやり直す必要をなくす点にある。もっとも、同制度の活用実績は限定的であり、特に疾病治療用 SaMD については、コアメカニズムの変更のみならずユーザーサポート機能の追加であっても臨床試験が求められる可能性が高いため、計画の策定・審査コストと手続の簡素化効果のバランスについて、実効性を疑問視する声もあがっている[20]。

---

[19] FDA, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[20] 厚生労働省「プログラム医療機器の特性を踏まえた適切かつ迅速な承認及び開発のためのガイダンス」第二版（令和 6 年 6 月 5 日）（https://www.pmda.go.jp/files/000269089.pdf）。IDATEN 制度の活用が実務上困難であることについては、「SaMD の不要な規制、開発意欲を削ぎかねず――

米国：FDA PCCP 制度

米国 FDA は 2024 年 12 月、AI 対応医療機器の PCCP（Predetermined Change Control Plan：事前確定変更管理計画）に関する最終ガイダンスを発行した[21]。PCCP では、製造業者が初回申請時に、①変更の範囲（Description of Modifications）、②変更を検証する方法論（Modification Protocol）、③変更の影響評価（Impact Assessment）を提示し、FDA がプロセス全体の妥当性を審査し承認する。承認範囲内の変更は追加申請不要で実装可能となる。さらに、PCCP は TPLC（Total Product Life Cycle）アプローチと一体的に運用される。すなわち、市販後のリアルワールドパフォーマンス（Real-World Performance: RWP）―実際の臨床現場における使用データに基づく性能評価―の継続的な監視が求められ、変更後の製品が想定どおりに機能しているかを市場で検証する仕組みが制度の不可欠な構成要素となっている。

PCCP の制度的意義を理解するには、それが単なる「変更手続の簡素化」ではないことに注意する必要がある。すなわち、PCCP においては、FDA の審査対象が「変更後の製品」から「変更を生み出すプロセスそのもの」へと転換されている。従来の変更審査では、変更のたびに FDA が変更後の製品を個別に審査していた。PCCP の下では、FDA は「変更をどのように検証し、どの基準で合否を判断するか」というプロセスの設計を審査し、そのプロセスが適切であると判断されれば、プロセスを通過した個々の変更結果については追加審査を要しない。具体的には、PCCP の審査では、製造業者が提出する変更プロトコルに含まれるデータの要件・検証手法・合格基準が、個別製品のリスク特性に照らして適切であるかが審査される（製品固有のリスク管理策の評価）。同時に、そのプロトコルを実行し継続的に遵守する組織の能力――すなわち品質マネジメントシステム、変更管理プロセス、市販後監視体制――もまた、FDA の審査対象となる（組織の変更管理能力の評価）。この二重の評価構造は、PCCP 単独で完結したものではなく、GMLP の策定（組織レベルの優良実践の標準化）や TPLC アプローチ（製品ライフサイクル全体を通じた管理）との連動の中で形成されてきた。

PCCP は、FDA が 2010 年代後半に試みた「Pre-Cert（事前認証）プログラム」の教訓を踏まえた制度である。Pre-Cert は、個別製品の審査に代えて開発組織の品質管理能力を認証し、認証を受けた組織の製品については市販前審査を省略するという野心的な構想であった。すなわち、「組織が優れていれば製品も安全」といえるかどうかを実証する試みだったのである。しかし、同プログラムは最終的

SaMD の現状と課題◆Vol.3」m3.com（2025 年 3 月 14 日）
（https://www.m3.com/news/open/iryoishin/1261551）参照。

[21] FDA, Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence/Machine Learning (AI/ML)-Enabled Device Software Functions: Final Guidance (December 2024).

に廃止された。その理由は複合的である。第一に、リスクが相対的に低いクラス 2 の医療機器であっても、組織の品質文化を KPI で評価するだけでは個別製品固有のリスクをカバーできないことが判明した。第二に、FDA が認証に必要な情報を企業から収集する法的権限を十分に有していなかった。第三に、パイロットプログラムの参加企業が Apple、Johnson & Johnson など大企業に偏り、スタートアップから不公平との批判を受けた [22]。

Pre-Cert の廃止が示す教訓は、組織認証のみでは製品安全を保証できないということである。しかし同時に、組織認証の手法が完全に無意味であったわけではない。Pre-Cert の部分的成果は、2021 年の FDA・英国 MHRA・カナダ Health Canada による「AI/ML 医療機器の開発組織に関する優良実践（Good Machine Learning Practice: GMLP）」の策定 [23]、さらには 2024 年の PCCP の最終制度化へとつながった。すなわち、現在の PCCP 制度は、組織の変更管理能力の事前評価と個別製品のリスク管理策の審査を組み合わせる方向に収斂した結果であり、「製品認証と組織的品質管理の段階的統合」という方向性を体現している。

米国：IDx-DR 自律型 AI 承認の先行事例

2018 年、FDA は糖尿病網膜症診断 AI「IDx-DR（現 LumineticsCore）」を、医師の判断を介さずに AI が単独で診断結果を出力することを認める形で、De Novo 経路――すなわち、既存の先行機器（predicate）が存在しない新規医療機器に対し、新たな規制分類を創設して許可する経路――により承認した [24]。これは、「特定領域において AI が人間の専門家なしに医療判断を行う」ことを規制当局が公式に認めた最初の事例であり、UHAI の社会実装に向けた先例として極めて重要な意義を持つ。

IDx-DR の承認構造には、UHAI の制度設計にとって示唆的な特徴がいくつかある。第一に、最終出力の責任は AI 開発企業が負う設計となっており、製品認証の枠組みの中で責任主体が明確に画定されている。第二に、開発企業だけでなく、使用施設に対してもトレーニング要件が設定されており、製品側のリスク管理策と利用組織側の体制整備を統合的に求める構造となっている。第三に、承認は特定の使用条件（プライマリケアの現場における糖尿病網膜症のスクリーニング）に限定されており、運行設計領域（ODD）の画定に類似する「適用範囲の限定」が安全性確保の前提条件となっている。

---

[22] FDA, The Software Precertification (Pre-Cert) Pilot Program: Tailored Total Product Lifecycle Approaches and Key Findings (September 2022).
[23] FDA, Health Canada & MHRA, Good Machine Learning Practice for Medical Device Development: Guiding Principles (October 2021).
[24] Abràmoff, M.D., et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, 1, 39. 前掲注 7 と同一文献。

3.1.2 三つの技術的課題への対応

以上の制度的枠組みを踏まえ、以下では日米の医療機器 AI 規制が、UHAI の三つの技術的課題―ブラックボックス性、動的変化、ゴール設定の困難―に対してどのような示唆を提供しているかを整理する。

3.1.2.1　ブラックボックス問題への対応

AI 医療機器の判断過程が不透明であることに対し、日米の制度はそれぞれ異なるアプローチでこの問題に取り組んでいるが、いずれも「AI の内部を直接解明する」ことを断念し、間接的な手法で安全性を担保しようとしている点では共通する。

① 製品レベルの対応：入出力検証による「振る舞いの評価」

3.1.1 で述べたとおり、日本の IDATEN 制度では、変更後のアルゴリズムを一度固定（ロック）した上で、臨床性能試験や非臨床性能試験を通じて入出力の関係を検証する。ブラックボックスの「中身」を解明するのではなく、「振る舞い」を検証するアプローチである。しかし、特に疾病治療用 SaMD においては評価コストが高く、システムの迅速な改良を阻害する要因となっている [25]。

米国の PCCP 制度は、この振る舞い検証をさらに一歩進め、個別の変更結果ではなく、変更を生み出すプロセスの妥当性を審査対象とする。再学習に用いるデータの要件、検証手法、合格基準といった「変更プロトコル」が事前に承認されていれば、そのプロセスを経た結果はブラックボックスであっても制度的に許容される。これは、ブラックボックス性を「解消」するのではなく制度内に「包摂」するための重要な視点転換である。

② 組織レベルの対応：製品評価の限界を補完する「組織の管理能力評価」

しかし、製品の入出力検証のみでは、ブラックボックスな AI の安全性を十分に担保できない。入出力検証はあくまで特定の条件下でのテスト結果に過ぎず、実運用中に遭遇するあらゆる状況を網羅することは原理的に不可能だからである。そこで注目されるのが、製品評価に加えて「その製品を開発・運用する組織の管理能力」を併せて評価する動きである。

3.1.1 で詳述したとおり、FDA の Pre-Cert プログラムは組織評価のみで製品安全を保証しようとして失敗したが、その教訓は PCCP 制度に活かされている。PCCP では、製品の変更プロトコルの審査（製品レベル）と、そのプロトコルを実行する

---

[25] 厚生労働省「プログラム医療機器の特性を踏まえた適切かつ迅速な承認及び開発のためのガイダンス」第二版（令和 6 年 6 月 5 日）。特に疾病治療用 SaMD に関する臨床試験要件については、同ガイダンス III-2 参照。

組織の品質マネジメントシステムの審査（組織レベル）が統合されている。IDx-DR の承認構造においても、製品の使用条件の限定に加えて使用施設へのトレーニング要件が課されており、製品と組織の双方を対象とする評価が制度に内包されている。このように、ブラックボックスな製品の挙動リスクを、製品の入出力検証と組織的ガバナンスの二段階で管理するアプローチが、日米を通じて形成されつつある。

3.1.2.2 動的変化への対応

継続的に学習・変化する AI の安全性をどう担保するかは、医療機器規制の中核的課題である。

日本の IDATEN 制度は、変更計画という柔軟な枠組みを導入しながらも、運用上はアルゴリズムを変更のたびにロックし、PMDA への届出後 30 日間の確認期間を経なければ実装できない。これは AI の連続的な進化を「段階的なコマ送り」に強制変換するものであり、継続学習の技術的利点を実質的に制限してしまうリスクがある。

米国の PCCP 制度は、TPLC（Total Product Life Cycle：製品ライフサイクル全体を通じた管理）の考え方に基づき、市販後のリアルワールドパフォーマンス（Real-World Performance: RWP）監視を安全確保の中核に据える。事前にプロトコルが承認されていれば、個々の変更について追加承認を要せず実装可能とすることで、動的変化を制度内に取り込む道を開いている。もっとも、PCCP の活用実績はなお限定的である。FDA の医療機器データベースによれば、PCCP 認可を受けた機器は累計 67 件であり、そのうち 59 件（88%）が直近 2 年間に集中している。2024 年に承認された 3,000 件超の 510(k)のうち、PCCP を含むものは 41 件（約 1%）にとどまる。この数字は、制度としての方向性が確立されつつある一方で、実務への浸透は緒に就いたばかりであることを示している [26]。

3.1.2.3　ゴール・評価手法設定の困難への対応

「何をもって安全とするか」の基準設定は、医療機器 AI 規制の根底にある問題である。

日本の IDATEN 制度は、既存の医療機器規制の枠組み（有効性・安全性）をそのまま適用しており、AI 固有の安全性評価方法論は確立されていない。確率的リスクを伴う AI に対し、規制当局は臨床試験という時間のかかるエビデンスを要求する傾向にあり、結果として制度が硬直化するおそれがある。

---

[26] Gardner Law, Streamlining Device Changes with Predetermined Change Control Plans (PCCPs) (2025), citing FDA Medical Device Database.

米国の PCCP 制度は、「性能基準と合格基準」を製造業者に事前提示させることで基準設定の問題に対処するが、その基準自体の妥当性評価は開発者の自己検証に大きく依存する。FDA は事後的なリアルワールドパフォーマンス（RWP）の監視でこれを補完するが、その評価手法が確立されていない領域では、結局その手法自体を開発者が自身で決定することになる。そうすると、結局は組織の信頼任せになってしまうリスクがある。

この「出口評価の設計困難」がもたらす帰結は、ドイツの Digitale Gesundheitsanwendungen（DiGA：デジタルヘルスアプリケーション）制度の経験にも端的に表れている。DiGA は、医師が処方できるアプリ型医療機器（例えば、糖尿病管理アプリや認知行動療法アプリ）について、暫定的に保険適用を認めた上で、一定期間内に臨床的有効性のエビデンスを提出させるという「ファストトラック」制度である。入口のハードルを下げてイノベーションを促進しつつ、出口（継続登録）で有効性を事後検証する設計であった。しかし、ファストトラックが適用された 68 品目のうち、所定の期間内に有効性を証明して継続登録に至ったのはわずか 12 件（20%未満）にとどまった[27]。入口の簡素化と迅速化だけでは不十分であり、市販後に信頼性のあるデータに基づいて AI の性能を継続的に検証する仕組み——すなわちポストマーケット監視の公共インフラ——がなければ、制度全体の信頼性が損なわれることを DiGA の経験は示している。

さらに、国際標準の統合にも困難がある。医療機器の品質マネジメントシステム規格である ISO 13485 の改定は 2030 年頃まで予定されておらず、AI 固有の規格（ISO/IEC 42001 等）との統合は当面行われない見込みである[28]。すなわち、医療機器としてのドメイン固有の認証要件と、AI システムとしての固有の認証要件を制度的に接続するための国際的枠組みは、なお発展途上にある。

3.1.3 小括

本節の分析から、日米の医療機器 AI 規制について以下の方向性と限界が明らかとなった。ブラックボックス性に対しては、日米とも内部構造の解明を断念し、入出力検証（日本のロック後臨床評価、米国のプロトコル審査）で対処する方向に収斂している。さらに、製品の振る舞い検証のみでは限界があるため、組織の管理能力を併せて評価する「二段階評価」の方向性が PCCP や IDx-DR の承認構造に

[27] Dahlhausen, F., et al. (2022). There's an App for That, but Nobody's Using It: Insights on Digital Health Application Evidence Requirements and Reimbursement in Germany. Digital Health, 8. BfArM の DiGA データベースも参照。

[28] ISO 13485:2016 は 2025 年の体系的レビューにおいて改定不要（confirm）と判断され、2030 年頃まで現行版が有効とされている。TÜV NORD, "ISO 13485:2016 reconfirmed – Valid until April 2030" (2025)参照。

見られる。ただし、Pre-Cert の失敗が示すように、組織評価のみでは製品安全を代替できないという教訓も共有されている。

動的変化に対しては、日本の IDATEN 制度がロック方式による「段階的コマ送り」を採る一方、米国の PCCP 制度は TPLC アプローチと RWP 監視により変更を制度内に包摂する道を開いている。ただし、PCCP の活用実績は全承認の約 1%に留まり、実務への浸透は緒に就いたばかりである。

ゴール・評価手法設定の困難に対しては、「何をもって安全とするか」の基準が未確立であることが両国に共通する根本的課題である。日本は臨床試験に依拠する保守的アプローチ、米国は開発者の自己検証に依存する柔軟なアプローチを採るが、いずれも AI 固有の安全性評価方法論を確立するには至っていない。

3.2 ＜事例 2＞ 自動運転システム（AV）の認証

自動運転領域は、UHAI において「人間の介在の排除」が最も顕著に現れる分野である。車両の制御が「運転支援（レベル 2）」から「完全自動運行（レベル 4 以上）」へと移行する際、システムは単なる道具ではなく、人間に代わる「運転主体」となる [29]。本節では、日本、英国、米国（カリフォルニア州）の規制モデルを分析し、各制度がブラックボックス性、動的変化、およびゴール・評価手法設定の困難という三つの課題にどのように向き合っているかを明らかにする。

3.2.1 制度の概要

日本：特定自動運行許可制度

2022 年道路交通法改正（2023 年 4 月施行）により、レベル 4 相当の「特定自動運行」に対する許可制度が創設された。事業者（特定自動運行実施者）は、運行設計領域（ODD: Operational Design Domain、自動運転システムが安全に機能できる地理的範囲・天候・速度等の条件）、遠隔監視体制、サイバーセキュリティ対策等を含む特定自動運行計画を都道府県公安委員会に提出し、許可を受ける。遠隔監視を行う特定自動運行主任者の配置、走行データ記録装置（DSSAD: Data Storage System for Automated Driving）の搭載が義務付けられる。

英国：Automated Vehicles Act 2024

2024 年 5 月成立の自動運転車法（AVA）は、認可自動運転事業体（ASDE: Authorized Self-Driving Entity，自動運転車の「運転主体」として法的責任を負う企業）という新たな責任主体を創設した [30]。自動運転モード中、車内の人間は運

[29] SAE International, J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (Revised April 2021).
[30] Automated Vehicles Act 2024, c. 10 (UK). 同法は 2024 年 5 月 20 日に国王裁可を受けた。

転操作に関する法的責任から免除され、ASDE がすべての義務を負う。ASDE として認可されるには、システムが「注意深く有能な運転者（Careful and Competent: C&C 基準）」と同等以上の安全性を維持できることを証明しなければならない。

英国の ASDE モデルにおいて特に注目すべきは、車両型式の認証（type approval）と組織としての認可（ASDE 認可）の二層構造を採用している点である。すなわち、製品としての車両の安全性評価と、その製品を継続的に運用・管理する組織の能力評価を統合的に求める枠組みとなっている [31]。ASDE として認可されるためには、安全性に関する技術的能力（自動運転システムの設計・テスト・継続的監視を行う能力）、適切な企業統治体制（安全性に関する意思決定プロセスの整備）、財務的健全性（損害賠償に対応しうる資力）、および情報開示体制（当局への誠実な報告を行う組織的仕組み）が求められる。さらに、ASDE は認可取得後も、システムの安全性を継続的に監視し、問題を検知した場合は速やかに当局に報告する義務を負う。

さらに、責任法制の観点から AVA 制度で注目すべきは、自動運転モード中の事故に関する民事責任の構造である。AVA は、自動運転車が事故を起こした場合、車内の人間は運転操作に関する法的責任を免除される一方、ASDE が「注意深く有能な運転者と同等以上」の安全基準を満たさなかった場合の責任を負う構造を採用した。これは、従来の自動車事故責任が「運転者」個人に帰属していた法的構造を根本的に転換し、安全性の保証義務を組織（ASDE）に一元化するものである。この責任の移転は、UHAI において「誰が最終的な責任主体となるか」という問題に対する一つの立法的回答である。

他方で、AVA は ASDE に対し、組織として適切な安全管理体制を構築・維持していたことが責任を軽減する根拠となりうる仕組みも内包している。ASDE が認可要件を充足し、誠実義務（Duty of Candor）を遵守し、継続的な安全性監視を実施していたにもかかわらず事故が生じた場合、これらの組織体制の充足が一定のセーフハーバーとして機能する可能性がある。これは、刑事法の文脈では「合理的な措置を講じていたこと」（due diligence defence）として構成されており、ASDE が安全確保のために合理的に期待される措置を講じていたことを立証できれば、刑事責任を免れうる。この仕組みは、単に責任を組織に移転するだけでなく、適切なガバナンス体制の構築自体にインセンティブを付与するものであり、いわばコンプライアンスが防御手段となる（compliance as defence[32]）構造を制度化したものといえる。

---

[31] Law Commission of England and Wales & Scottish Law Commission, Automated Vehicles: Joint Report, Law Com No 404, Scot Law Com No 258 (January 2022).
[32] Automated Vehicles Act 2024, Part 1, section 10; Law Commission Joint Report (2022) paras 11.28 and following .

カリフォルニア州（米）：配備許可制度

カリフォルニア州車両管理局（DMV: Department of Motor Vehicles）は、段階的な許可制度（テスト許可→無人テスト許可→配備許可）を運用する[33]。連邦レベルでは米国運輸省道路交通安全局（NHTSA: National Highway Traffic Safety Administration）の自己認証制度の下、製造者が VSSA（Voluntary Safety Self-Assessment：任意の安全性自己評価報告書）を提出する。具体的な技術基準の策定は限定的であり、データ開示義務と事後監視を軸とする。

3.2.2 三つの技術的課題への対応

3.2.2.1 ブラックボックス問題への対応

自動運転 AI の判断過程が不透明であることに対し、三カ国はそれぞれ異なるアプローチを採る。

日本は、DSSAD による走行データの記録を義務付けるとともに、車両の安全基準の策定においては「注意深く有能な運転者（competent and careful human driver）」と同等以上の安全水準を達成しているかをシナリオベースで評価する方法論を開発している（SAKURA プロジェクト）。[34]AI の内部判断過程を直接検証するのではなく、結果（ログ及びシナリオテストの結果）に基づいて安全性を判定するアプローチである。ただし、このアプローチは結果のみに依拠しており、AI の内部論理を直接評価する手段を欠く。その結果として、遠隔監視を行う特定自動運行主任者を配置し、AI の不確実性を人間の介在によって補完する（すなわち、UHAI には至らない）制度設計である。

英国は、ASDE 制度により、ブラックボックス問題を「法的擬制」で回避する。AI が「なぜ」その判断に至ったかを問う代わりに、運転結果そのものを「ASDE の行為」として帰属させる。これにより、AI 内部が不透明であっても、C&C 基準を逸脱した場合には ASDE が法定な責任を負う。技術的なブラックボックスを組織責任に読み替えることで、明確な法的回答を提示したものといえる。そのため、ASDE が安全性の保証を継続的に担えるよう厳格な要件が課されている。

カリフォルニア州は、VSSA による自己評価と事故データの開示義務を通じて透明性を確保する。内部判断過程の解明よりも、走行実績データの公開に重点を置く。

[33] Cal. Veh. Code §§ 38750–38755; Cal. Code Regs. tit. 13, §§ 227–228.
[34] SAKURA プロジェクト。https://www.sakura-prj.go.jp/jp/

3.2.2.2 動的変化への対応

OTA（Over-the-Air：無線通信を通じたソフトウェア遠隔更新）アップデートにより継続的に変化するシステムの管理は、自動運転規制の核心的課題である。

日本の制度は、特定のソフトウェアバージョンと ODD の組み合わせに基づくスナップショット評価を前提とする。アップデートのたびに変更届出や再審査が必要となり得るため、AI の動的な進化と行政許可の静的な性格との間に構造的な課題が生じている。

英国は、ASDE に対する「誠実義務（Duty of Candor）」――すなわち、安全上のリスクや不具合を認知した場合に、これを隠蔽せず速やかに当局に報告する義務――と継続的安全性監視を軸とする動的な管理モデルを採用する。特筆すべきは、企業内で情報提供責任者となるシニアマネージャーを指名させ、虚偽情報の提供や意図的な隠蔽に対しては個人に刑事責任を課す点である。この制度設計は、企業と個人の間に構造的な緊張関係を意図的に創出するものである。仮に企業が安全上の問題を組織的に隠蔽しようとしても、指名されたシニアマネージャー個人が刑事責任を問われるリスクを負うため、個人にとっては隠蔽に加担するよりも当局に報告するインセンティブが働く。この「企業の隠蔽利益」と「個人の免責利益」の不一致は、いわば囚人のジレンマに類似する構造であり、情報開示を促進するための制度的担保として機能することが意図されている。また、当局は運行データに基づき認可の変更・停止・取消を機動的に行う権限を有する。これは、事前の一時的な認可に依存せず、運用段階のパフォーマンスを重視する動的な規制モデルである。

カリフォルニア州は、製造者の自己認証に基づくため、更新は原則として開発者の判断で行われる。DMV は配備許可の条件として運用段階の監視権限を留保し、事故報告・走行障害報告等のデータ開示義務を通じて事後的な安全確保を図る。この事後監視モデルが実際に執行された事例として、2023 年 10 月の GM Cruise[35] 社に対する配備許可の停止がある。同社の無人自動運転タクシーが歩行者を巻き込む事故を起こした際、Cruise 社が事故の全容―特に、車両が歩行者を引きずった事実―を DMV に対して速やかに開示しなかったことが問題視された。DMV は配備許可を停止し、同社は全米での無人運行を中断した。この事例は、データ開示義務に基づく事後監視モデルが、開示が適切に行われなかった場合には機能不全に陥りうること、そして規制当局が許可停止という強力な手段を行使しうることの双方を示している。英国の Duty of Candor と個人刑事責任の組み合わせは、

[35] California DMV, Order of Suspension of Driverless Deployment Permit, Cruise LLC (October 24, 2023).

こうした情報隠蔽を事前に抑止する制度設計として、Cruise 社の事例が浮き彫りにした課題への一つの回答と位置づけられる。

3.2.2.3 ゴール・評価手法設定の困難への対応

「安全」の定義をどう設定するかは、三カ国に共通する根本的課題である。英国は AVA において C&C 基準（注意深く有能な運転者）を明示的な法的ベンチマークとして採用し、日本も WP.29 の FRAV（Framework document on automated/autonomous vehicles）での議論を踏まえ、車両安全基準の策定において C&C 基準を参照している（SAKURA プロジェクト及び国土交通省「自動運転車の安全性能確保措置に関する検討会」最終報告書（2025 年 3 月））。[36]米国はこのような統一的な定性基準を公式には採用しておらず、開発者による自己評価（VSSA）に委ねる構造となっている。しかし、いずれの国も「AI の安全性をどのような基準で測定・判断すべきか」という根本問題に直面しており、この定性的基準を実装可能な形に翻訳する際に共通の困難が存在する。

第一に、「意味論的ギャップ」の問題がある。法が求める「注意深さ」という抽象的な概念を、エンジニアがプログラム可能で検証可能な技術仕様へと翻訳するプロセスは非常に困難である。日本の特定自動運行許可制度では、ODD の定義を通じて運行条件を限定することで間接的にこのギャップを緩和しているが、C&C 基準そのものを技術仕様に翻訳する方法論には着手したばかりである。英国は Statement of Safety Principles（SoSP）の策定を通じ、C&C 基準をより具体的な安全原則に分解する試みを進めているが、これはなお法的拘束力を持つ基準ではなく、技術仕様への最終的な「翻訳」は開発者に委ねられる構造となっている [37]。米国では、開発者が VSSA を通じて自らの安全性アプローチを説明する方式を採るが、その妥当性を検証する統一的な基準は存在しない。

第二に、法的予見可能性の問題がある。基準が定性的である限り、事故発生後に裁判所が「後知恵バイアス（Hindsight Bias）」に基づいて安全基準を再定義するリスクを排除できない。この問題は、セーフハーバーの範囲と限界をいかに明確に画定するかという論点に直結する。

第三に、安全性の評価手法自体の問題がある。統計的な事故率のみでは安全性を十分に評価できないことは認識されているものの、形式手法（Formal Methods）や実行時モニタリング（Runtime Monitoring）といった代替的評価手法を現行法の

[36] 国土交通省「自動運転車の安全性能確保措置に関する検討会」最終報告書（2025 年 3 月 3 日）。https://www.mlit.go.jp/jidosha/content/001879642.pdf
[37] Department for Transport (UK), Statement of Safety Principles for Automated Vehicles: Consultation Document (2025).

「注意義務」の枠組みにどう位置づけるかは、いずれの制度においても未確立である。

3.2.3 小括

自動運転の規制制度は、従来の「装置の安全審査」から「組織・プロセス・行動のガバナンス」へとパラダイムシフトの渦中にある。ブラックボックス性に対しては、三カ国とも内部判断過程の解明を断念し、間接的な手法で対処している。日本は DSSAD のログ記録とシナリオベース評価、英国は ASDE への法的擬制（結果を組織に帰属させる）、米国は走行実績データの開示義務をそれぞれ軸とする。いずれも「振る舞い」の事後検証に依拠しており、英国の ASDE モデルがこれに組織ガバナンスの評価を統合している点が特徴的である。

動的変化に対しては、日本のスナップショット評価（バージョン固定＋変更届出）と、英国の動的管理モデル（Duty of Candor＋個人刑事責任＋認可の機動的変更・停止）が対照的である。米国は開発者の自己判断に委ね、事後的なデータ開示で補完する。実際に、Cruise 社は不適切な情報開示を理由の一つとして営業許可を取り消されることとなった。

ゴール設定の困難に対しては、英国が C&C 基準を明示的に採用し日本もこれを参照する一方、米国は統一基準を欠く。いずれの国においても、定性的な安全基準を検証可能な技術仕様に翻訳する方法論は未確立であり、「意味論的ギャップ」が法的予見可能性の確保を阻んでいる。

3.3 ケーススタディの総括：二つの領域から導かれる制度的含意

本章では、医療機器 AI（3.1）および自動運転（3.2）の二領域を対象に、主要国の規制制度が UHAI の三つの技術的課題——ブラックボックス性、動的変化、ゴール・評価手法設定の困難——にどのように向き合っているかを分析した。本節では、これらの知見を第 2 章で提起した三つの制度的課題——「評価」「認証」「法制度」——の文脈で整理する。

3.3.1 評価に関する知見

両領域の規制制度は、AI の内部構造の解明を直接求めることが不可能であることを前提として、入出力の検証を通じた安全性評価——すなわち「振る舞いの検証」——に収斂している。医療機器においては IDATEN 制度のロック後評価や PCCP のプロトコル審査が、自動運転においては DSSAD のログベース評価やシナリオベースの C&C 基準が、その具体的な現れである。これは、ブラックボックス性に対して、内部の解明ではなく外部からの検証で対処するという共通の方向性を示している。

しかし、この入出力評価の手法――AI 固有のリスクに対応する体系的な V&V（検証・妥当性確認）――は確立されていない。そのような製品レベルの評価手法の限界を補うものとして、両領域で共通して観察されるのが、製品評価に加えて組織の管理能力を併せて評価する動きである。3.1 で分析した PCCP の二重評価構造（変更プロトコルの適切性＋組織の変更管理能力）や、3.2 で分析した英国 ASDE の二層構造（車両型式認証＋組織認可）は、ブラックボックスな製品の安全性を、製品単体の試験結果だけでなく、その製品を開発・運用する組織のガバナンス体制との組み合わせで担保しようとする共通の方向性を示している。

動的変化に対しては、スナップショット的な一時点評価の限界をどう克服するかが問題となる。米国の PCCP 制度の TPLC アプローチ（製品ライフサイクル全体を通じた管理）や英国の Duty of Candor は、ライフサイクル全体を通じた継続的評価の方向性を示すが、いずれも実務への浸透は緒に就いたばかりである。ドイツの DiGA の経験は、入口の簡素化だけでは不十分であり、ポストマーケット監視のインフラ整備が制度の信頼性に不可欠であることを裏付けている。

ゴール設定の困難については、「何をもって安全とするか」についての社会的合意を形成する制度的メカニズムが未確立である。自動運転の文脈では、「人間の専門家と比較して同等以上」という比較基準の方向性は、共有されているものの、その統計的な閾値の設定方法や、「公平性」「説明可能性」といった定性的要件の測定方法は確立されていない。

これらの知見は、V&V 手法の標準化（レッドチーミング、メタモルフィック・テスティング、形式手法等（後述の 4.1.2 参照）の複合的活用）、ライフサイクル全体を通じた継続的モニタリングの制度化、およびリスク許容水準の社会的合意形成の必要性を示唆している。

3.3.2 認証に関する知見

両領域の分析から浮かび上がる重要な教訓は、製品認証のみでは AI のような動的システムの安全性を継続的に担保することは困難である一方で、組織認証のみでも製品安全を保証できないということである。

現行制度は、この認識に基づき、製品評価と組織評価の統合に向かいつつある。米国の PCCP は、製品固有のリスク管理策の審査と、組織の変更管理能力の事前評価を組み合わせている。英国の ASDE モデルは車両型認証と組織認可の二層構造を採用している。IDx-DR の承認構造――開発企業の責任と使用施設のトレーニング要件の統合――もまた、この方向性の先行事例である。

その上で、認証における核心的課題は、規制当局や認証機関がブラックボックスな AI システムの安全性をどこまで独立に評価しうるかという、能力と方法論の不

足にある。医療機器領域では、PMDA や FDA は臨床データの審査能力を有するが、AI アルゴリズム自体の技術的評価については専門性が限定的である。自動運転領域では、日本の都道府県公安委員会は AI システムの技術的評価能力に限界があり、米国の NHTSA は製造者の自己認証に依拠している。英国でさえ、VCA（Vehicle Certification Agency）が AI アルゴリズムの安全性を直接評価する方法論は発展途上にある。すなわち、「製品と組織の統合的評価」という方向性が見えつつある一方で、その評価を実施する主体の能力構築が追いついていないのが現状である。

さらに、ドメイン固有の認証（ISO 13485（医療機器の品質マネジメントシステム規格）、車両型式認証等）と AI 固有の認証（ISO/IEC 42001 等）を制度的に接続するための国際的枠組みも発展途上にある。ISO 13485 の改定は 2030 年頃まで予定されておらず、当面は両者の制度的統合は見込まれない。

これらの知見は、製品のリスク管理策と組織のマネジメント能力を統合的に評価する「ジョイント認証」の制度設計、認証機関の技術的能力の構築、および既存のドメイン認証との効率的な接続と国際的な相互運用性の確保の必要性を示唆している。

3.3.3 法制度に関する知見

両領域のケーススタディからは、法制度に関して以下の示唆が導かれる。

第一に、ブラックボックス性への法的対応として、各国は AI 内部の因果関係の追及を断念する方向に向かっている。英国 ASDE の法的擬制はその最も明確な表現であり、技術的なブラックボックスを組織責任に読み替えることで、因果関係の追及を迂回している。PCCP の「プロトコル遵守」が将来的にセーフハーバーとして機能する可能性、IDx-DR の De Novo 承認が AI の自律的判断の法的受容の先例として持つ意義は、いずれもブラックボックス性と許認可制度・責任法制の接合点を探る上で重要な素材である。さらに注目すべきは、英国 AVA が導入した「コンプライアンスが防御となる（compliance as defence）」構造である。ASDE が認可要件を充足し、誠実義務を遵守し、合理的な安全管理措置を講じていた場合には、due diligence defence として刑事責任を免れうる仕組みが整備されている。これは、ブラックボックスな AI の内部因果を解明できなくとも、組織的ガバナンス体制の充足自体が一定のセーフハーバーとして機能しうることを制度的に示したものであり、UHAI における責任法制の設計にとって重要な参照点となる。

第二に、動的変化への法的対応として、事前の一時点評価のみでは不十分であり、製品のライフサイクル全体を通じた継続的管理が不可欠であるという認識が各国に共有されつつある。しかし、継続的に変化する AI システムの状態を外部の規制当局が常時直接監視することは、技術的にも資源的にも現実的ではない。そ

こで各国の制度が収斂しつつある方向性は、継続的管理の第一義的な責任を事業者自身に課しつつ、その自主的管理を実効化するための制度的担保を設けるというものである。英国の Duty of Candor は、事業者に安全上の問題の自主的報告を義務付け、虚偽報告や隠蔽に対してはシニアマネージャー個人に刑事責任を課すことで、自主報告のインセンティブを構造的に確保している。米国の PCCP における RWP 監視義務も同様に、市販後の変化を事業者が自ら継続的に検証する責任を制度化したものである。すなわち、動的変化への法的対応の要諦は、「外部からの全面的監督」ではなく「事業者の自主的監督＋報告義務＋違反時の制裁」の組み合わせによる実効性の確保にある。

第三に、ゴール設定の困難への法的対応として、各国の制度は「何をもって安全とするか」の合意形成を、製品性能・テスト方法論・組織体制を統合的に対象とする枠組みとして構築する方向に向かいつつある。C&C 基準に代表される定性的安全基準の「意味論的ギャップ」は、単に技術仕様の翻訳困難という問題にとどまらず、事前の法的予見可能性―すなわち、事業者がどこまで対応すれば法的責任を免れうるかの見通し―を確保できないという問題に直結する。この不確定性は、事故発生後の後知恵バイアスのリスクも内包する。こうした課題に対し、PCCP における「性能基準＋検証手法＋合格基準」の事前承認、英国の SoSP（安全原則声明）による C&C 基準の具体化、そして ASDE の due diligence defence は、いずれも「製品がどのような性能を示すべきか」「それをどのようなテスト環境で検証すべきか」「組織としてどのような管理体制を整備すべきか」という三つの次元を統合的にセーフハーバーとして画定する試みと理解できる。UHAI の文脈においても、セーフハーバーの設計は、製品・テスト・組織のいずれか単独ではなく、三者を統合した形で行われる必要があることを、各国の制度的試みは示唆している。

本章で検討した各国の制度的対応は、UHAI に対する部分的な解を示すのものであるが、いずれの制度もまだ発展途上にある。そこで、第 4 章では、これらの制度的工夫を参照しつつ、UHAI の実装を後押しするための評価基準（4.1）、認証の仕組み（4.2）、および法制度（4.3）がそれぞれ独立に機能し、かつ相互に補完しあう枠組みを構想する。

【参考事例】宇宙システムにおける異常予兆検知と AI 安全性評価

本章で検討した医療機器 AI および自動運転とは異なる領域であるが、宇宙システムの運用においても、三つの本質的課題への対応が模索されている。国際宇宙ステーション（ISS）日本実験モジュール「きぼう」の熱制御システム（TCA-

L）に関する研究（Iino et al.）では、複数パラメータの相互依存関係を考慮した異常予兆検知手法を構築し、2012 年に発生した結露に起因するポンプインバーターの不具合について、正常挙動データに基づく予測モデルと実測値の誤差増加から事前に異常の兆候を検出するとともに、SpecTRM-RL（形式手法）と呼ばれる形式手法を組み合わせることで、同予兆の検知に用いたデータの特徴から、異常の原因まで示す手法が提案されている [38]。

この研究が示唆するのは、三つの課題に対する一つの実践的アプローチである。第一に、ブラックボックス問題に対しては、FRAM（機能共鳴分析手法）による重要パラメータの体系的選定と SpecTRM-RL（形式手法）による事後検証を組み合わせることで、AI の判断に対する説明可能性（Explainability）を確保している。具体的には、SpecTRM-RL により正常時と異常時のパラメータ条件の組み合わせを比較することで、結露がポンプインバーター故障の原因であることを特定でき、その結果は当時実際に行われたトラブルシューティング結果とも一致した。第二に、動的変化に対しては、ランダムフォレストによる正常挙動モデルを用いて予測誤差を継続的に監視し、環境変化に応じて FRAM モデルの更新とパラメータ追加を行う適応的な枠組みを採用している。第三に、評価基準の設定に関しては、単純な閾値超過によるアラートではなく、複数パラメータ間の相互依存関係をシステム理論的に把握した上で統計的閾値を設定するアプローチを採っており、モデル選定にも Pugh Concept Selection により検出精度・検出速度・予測精度の多次元的評価を行っている。

「きぼう」は地上約 400km の軌道上にあり、筑波宇宙センターから 24 時間 365 日の常時監視が行われているが、その際に地上の管制官は、数百に及ぶテレメトリデータから異常の兆候も含めた状況を複合的に判断しなければならない。そのスキルは管制官個人の技量や経験に依存する面も多く、そのような複雑なタスクを機械やコンピュータにより支援するための要求も、昨今の社会情勢の中で増しつつある。この環境は、通信遅延と物理的距離により人間の即時介入が制約されるという点で、UHAI が提起する「人間介在の前提が成立しにくい環境」の一つの先行事例と位置づけることができる。同研究が示す「説明可能な AI 監視＋形式手法による検証」の組み合わせは、本章で検討した各国制度が共

[38] Iino, S., Nomoto, H., Fukui, T., Ishizawa, S., Sasaki, M., Yagisawa, Y., Hirose, T., Michiura, Y. & Shibayama, H. (2023). Systemic symptom detection in telemetry of ISS with explainability using FRAM and SpecTRM. Proceedings of 4th Asia Pacific Conference of the Prognostics and Health Management Society (PHMAP 2023), OS02-03; Iino, S., Nomoto, H., Fukui, T., Yagisawa, Y., Ishizawa, S., Hirose, T. & Michiura, Y. (2024). Towards Explainable Anomaly Detection in Safety-critical Systems Employing FRAM and SpecTRM in International Space Station Telemetry. International Journal of Prognostics and Health Management, 15(3). doi:10.36001/ijphm.2024.v15i3.3857.

通して直面するプロセス評価基準の具体化に向けた技術的方向性を示唆するものである。

4. 今後の制度構築に向けた論点整理

第 2 章では、UHAI の三つの技術的特性——ブラックボックス性、動的変化、ゴール・評価手法設定の困難——が、「評価」「認証」「法制度」の三層にわたって構造的な課題を生じさせることを論じた。第 3 章のケーススタディでは、医療機器 AI および自動運転の各国規制がこれらの課題に部分的に対処しつつも、いずれも人間の介在を何らかの形で前提としており、UHAI に対しては不十分であることを確認した。

本章では、これらの分析を踏まえ、UHAI の社会実装を可能にするための制度構築において今後議論すべき論点を、「評価」（4.1）、「認証」（4.2）、「法制度」（4.3）の三層に即して整理する。その上で 4.4 において、三層が相互にどう連関し補完しあうべきかを論じ、第 2 章で指摘した「悪循環」を断ち切るための視座を提示する。

4.1 UHAI の評価をめぐる論点

4.1.1 リスク許容水準の社会的合意

UHAI の評価における最も根本的な問いは、「何をもって安全とするか」である。AI は確率的に挙動するシステムである以上、ゼロリスクを保証することは原理的に不可能であり、「どの程度のリスクであれば社会的に許容できるか」という基準を定義することが、あらゆる評価の出発点となる。

第 3 章で確認したとおり、英国の C&C 基準（注意深く有能な運転者）や日本の SAKURA プロジェクトは、「人間の専門家と比較して同等以上」という比較基準の方向性を示している。この方向性は、絶対的な安全基準の設定が困難である UHAI にとって、有力な選択肢である。ただし、ここでいう「同等以上」とは、個々の判断がすべて人間の判断と一致することを意味しない。個別の判断において AI が人間とは異なる誤りを犯しうる可能性は排除されないが、結果全体として統計的に評価した場合に、適切な能力を有する人間の集団と比較して同等以上の安全性が確保されていることが、許容の条件となる。すなわち、問われるべきは個々の挙動の同一性ではなく、総体としてのリスク水準の比較である。

しかし、この比較基準にも検討すべき課題がある。第一に、「適切な能力を有する人間」の水準自体が領域によって大きく異なるため、比較のベースラインを一律に設定することが困難である。第二に、「同等以上」であるかを統計的に検証するための方法論——サンプルサイズ、比較条件の設定、有意差の判断基準——が確立されていない。第三に、物理的な安全性に加えて「公平性」や「説明可能性」といった定性的要件をどのように測定・評価すべきかという方法論自体が未確立である。

さらに、リスク許容水準の決定は純粋に技術的な判断ではなく、倫理的かつ政治的な判断を含む。どのようなリスクをどこまで受け入れるかは、最終的には社会的選択の問題である。したがって、政府、事業者、アカデミアに加え、消費者団体や被害者団体を含む多様なステークホルダーが参加する透明性の高いプロセスで安全目標を策定する必要がある。この合意形成プロセスをどのように制度化するか——官民のどちらが議論をリードするのか、どのような領域ごとに設計すべきか、国際的な調和をどう図るか——は、評価基準の構築における最も重要な論点の一つである。

4.1.2 製品評価と組織評価の統合

第 3 章の分析は、製品の入出力検証（「振る舞いの検証」）のみではブラックボックスな AI の安全性を十分に担保できないことを示した。入出力検証はあくまで特定のテスト条件下での結果に過ぎず、実運用中に遭遇するあらゆる状況を網羅することは原理的に不可能だからである。

この限界に対しては、まず製品レベルの評価手法を高度化する方向性がある。従来の医療機器や自動車の安全試験は「仕様どおりに動作するか」を検証するものであったが、UHAI では「仕様」自体が確定できないため、検証の対象を再定義する必要がある。具体的には、入出力の関係性が一定の条件下で安定しているかを検証するメタモルフィック・テスティング、安全な動作範囲からの逸脱がないことを数学的に証明する形式手法（ニューラルネットワーク検証）、意図的に敵対的入力を与えて脆弱性を発見するレッドチーミング等の手法が注目されている。[39] 第 3 章の参考事例で紹介した宇宙システム「きぼう」における異常予兆検知研究は、形式手法（SpecTRM-RL）を用いることで、人間の即時介入が制約される環境

[39] メタモルフィック・テスティングについては、Segura, S., Fraser, G., Sanchez, A.B. & Ruiz-Cortés, A. (2016). A Survey on Metamorphic Testing. IEEE Transactions on Software Engineering, 42(9), 805‑824 参照。

ニューラルネットワーク検証の形式手法に関するサーベイとして、Meng, W., et al. (2025). Adversarial Robustness of Deep Neural Networks:A Survey from a Formal Verification Perspective. IEEE Transactions on Dependable and Secure Computing, 22(1), 243‑264 参照。

においても AI の判断に対する説明可能性を技術的に担保できることを示しており、UHAI の V&V 手法の方向性を考える上で示唆的である。ただし、これらの手法を認証プロセスにどのように組み込み、どの水準の検証をもって「十分」とするかは、V&V の標準化における核心的な問いとして残されている。

しかし、こうした製品レベルの評価手法がいかに高度化されても、テスト条件下の安全性しか保証できないという本質的な限界は克服しがたい。この限界を補完するものとして、3.3.1 で確認したとおり、製品評価に加えて「その製品を開発・運用する組織の管理能力」を併せて評価する方向性が、米国の医療機器規制（PCCP の二重評価構造）および英国の自動運転規制（ASDE の二層構造）に共通して見られる。

ここで検討すべき論点は、製品評価と組織評価の比重配分をどう設計すべきかという問題である。組織評価の比重を高めすぎれば個別製品固有のリスクが見落とされる。他方で、製品の入出力評価に固執すれば、スナップショット評価の限界から逃れられず、AI の動的な性質に対応できない。PCCP が到達した「変更プロトコルの適切性（製品レベル）＋プロトコル実行能力（組織レベル）」という二重構造は、両者のバランスを取る一つの方向性を示しているが、このバランスの最適解はリスクの性質や領域によって異なりうる。どのような場合に製品評価を重視し、どのような場合に組織評価に依拠しうるかの基準を策定することが、今後の検討課題となる。

4.1.3 ライフサイクルを通じた継続的評価

動的変化に対するスナップショット評価の限界は、第 3 章で繰り返し確認された。したがって、UHAI の評価は、出荷前の一時点評価と市販後の継続的監視を一体的に設計する必要がある。論点は以下の二つに整理される。

第一に、ポストマーケット監視の制度化である。PCCP における RWP（リアルワールドパフォーマンス）監視や英国の Duty of Candor に見られるように、市販後の継続的な安全性検証を制度に組み込む方向性は明らかであるが、その具体的な設計——監視の頻度、指標、閾値、報告義務の範囲——は確立されていない。さらに、監視を個々の事業者に委ねるだけでは不十分であり、事故やヒヤリハット情報を業界横断的に共有・分析する公共インフラの整備が検討されるべきである。OECD の「グローバル AI インシデント報告フレームワーク」[40]のような国際的な情報共有の枠組みとの連携も論点となる。

[40] OECD, Towards a Common Reporting Framework for AI Incidents (2024). https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html

第二に、出荷前評価と市販後監視の制度的接続である。現行制度では、出荷前の許認可と市販後の監視は制度的に分離されていることが多い。しかし UHAI においては、市販後の監視結果が許認可の維持・変更・取消にどのように連動すべきかが重要な設計課題となる。運用中に AI が安全な動作範囲を逸脱したかを常時監視する実行時モニタリング（Runtime Monitoring）――逸脱を検知した場合には安全なフォールバック制御へ移行させる「安全エンベロープ」の仕組み[41]――をポストマーケット監視と許認可制度にどう連動させるかは、動的許認可の制度設計にとって核心的な論点である。

4.2 UHAI の認証をめぐる論点

4.2.1 製品認証と組織認証の統合――ジョイント認証の可能性

3.3.2 で確認したとおり、Pre-Cert の失敗（組織認証のみでは不十分）と、PCCP の二重評価構造（製品＋組織）の形成は、製品認証と組織認証のいずれか一方では足りず、両者を統合する必要性を実証的に示している。

この方向性を制度化する一つの構想として、「ジョイント認証（Joint Certification）」が考えられる。ジョイント認証とは、AI システム固有のリスクに対する具体的な管理策（Controls）――バイアス対策、テスト手法、安全エンベロープ等――の適切性と、それを継続的に運用・改善する組織のマネジメント能力――品質マネジメントシステム、変更管理プロセス、インシデント対応体制等――をセットで評価する仕組みである。国際標準の文脈では、ISO/IEC 42001（AI マネジメントシステム）の本文（組織のマネジメント要件）と附属書 A/B（管理策）の構造が、この統合的評価の基盤を提供しうる。[42]

ジョイント認証の構想に関しては、いくつかの論点がある。第一に、製品固有のリスク管理策の審査において、4.1 節で論じた V&V 手法（レッドチーミング、形式手法、実行時モニタリング等）の導入状況をどのように評価指標とすべきかである。第二に、ジョイント認証がリスクレベルに応じた多層的な制度として設計されるべきであること――すべての AI に同一の厳格な認証を求めるのではなく、

[41] 実行時モニタリング（Runtime Monitoring / Runtime Verification）の概念と自律システムへの適用については、Leucker, M. & Schallhart, C. (2009). A Brief Account of Runtime Verification. Journal of Logic and Algebraic Programming, 78(5), 293-303 参照。自動運転における安全エンベロープ（Safe Driving Envelope）の検証については、Mehmed, A. (2021). Runtime Monitoring for Safe Automated Driving Systems. Doctoral Thesis, Chalmers University of Technology 参照。

[42] ISO/IEC 42001:2023, Information technology — Artificial intelligence —Management system. 同規格は、AI マネジメントシステムに関する世界初の国際規格であり、組織のマネジメント要件（本文 Clauses 4-10）と、AI 固有の管理策（Annex A）及びその実施ガイダンス（Annex B）から構成される。

リスクの大きさに応じて求められる認証の深度と範囲を段階的に設定すべきこと——である。第三に、組織認証を一度取得すれば、同一組織が提供する複数の AI システムについては製品固有のリスク管理策のみを追加審査すれば足りるという効率化の可能性である。この「組織認証の流用」は、認証コストの削減とイノベーション促進の観点から重要であるが、組織認証の有効期間や更新要件の設計と併せて検討される必要がある。

事実、国際標準や各国の枠組みにおいても、この統合的アプローチの萌芽は既に見られる。EU AI 法では、高リスク AI システムの提供者に対し、製品のリスクマネジメントやデータガバナンス（管理策要件）と品質マネジメントシステム（組織要件）の双方の適合性評価を義務付けている[43]。米国 NIST の AI Risk Management Framework（AI RMF[44]）も、組織のガバナンス体制（Govern）を土台として個々のシステムのリスク管理（Map, Measure, Manage）を実践する構造を採る。日本国内でも、AI ではなくクラウドの政府調達要件に関するものではあるが、「政府情報システムのためのセキュリティ評価制度（ISMAP）」が組織的マネジメント体制と個別システムの管理策の双方を求めるアプローチを採っており、ジョイント認証の方向性が実務的に実現可能であることを示唆している。[45]

4.2.2 認証機関の能力構築

ジョイント認証の制度設計にとって、それ以上に本質的な課題は、その認証を誰が実施しうるのかという問いである。3.3.2 で確認したとおり、規制当局や認証機関がブラックボックスな AI システムの安全性を独立に評価する能力は、いずれの国においても限定的である。

医療機器領域では、PMDA や FDA は臨床データの審査能力を有するが、AI アルゴリズム自体の技術的評価——例えば、学習データのバイアス評価、モデルのロバストネス検証、ドメイン外での性能劣化の予測——については専門性が十分ではない。自動運転領域では、日本の都道府県公安委員会は AI システムの技術的評価能力に構造的な限界があり、米国の NHTSA は製造者の自己認証に大きく依拠している。英国でさえ、VCA（Vehicle Certification Agency）が AI アルゴリズムの安全性を直接評価する方法論は発展途上にある。

この能力ギャップを埋めるためには、以下の論点が検討されるべきである。第一に、認証機関における AI 技術の専門人材の育成と確保である。第二に、認証機関間の国際的な知見共有——ある国の認証機関が開発した評価手法を他国の機関が活用できる枠組み——の制度化である。第三に、認証プロセスへの民間専門機関

[43] Regulation (EU) 2024/1689, Articles 9, 17, 40–49.
[44] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (January 2023).
[45] 内閣サイバーセキュリティセンター（NISC）「政府情報システムのためのセキュリティ評価制度（ISMAP）」。https://www.ismap.go.jp/

の活用——官民連携による審査体制の構築——の可能性とその条件である。ジョイント認証の「形」を設計しても、それを実施できる主体の能力が伴わなければ制度は機能しない。能力構築は認証制度設計の不可欠な前提条件であり、制度設計と並行して進められるべきである。

4.2.3 既存ドメイン認証との接続と国際的相互運用性

UHAI は、医療機器、自動車、金融システム等の既存の規制ドメインにおいて実装される。したがって、AI 固有の認証要件をゼロから構築するのではなく、各ドメインに既に存在する認証制度との効率的な接続が不可欠である。

例えば、医療機器の品質マネジメントシステム規格である ISO 13485 と AI の ISO/IEC 42001 は、PDCA サイクルに基づくマネジメントシステムという共通構造を持つ。ISO 13485 を満たす医療機器メーカーが、AI 特有の追加的要求事項（AI システム影響評価、データガバナンス等）のみをクリアすることでジョイント認証を取得できるよう制度を接続すれば、重複審査を避ける効率的な運用が実現しうる。しかし、これらの実現は、既存規格の改版や新規規格の作成が必要であり、AI 固有の規格との制度的統合は短期間で実現することは見込まれない。

この状況を踏まえると、短中期的には国際標準の統合を待たずに、各ドメインの認証制度と AI 認証の「接続インターフェース」を設計する実務的なアプローチが求められる。ISO/IEC 42007[46]等で AI システムの適合性評価スキームの国際標準化が進められており、日本の認証制度もこれとの整合性を保つことが、グローバルな相互運用性の確保にとって重要である。

4.3 UHAI の法制度をめぐる論点

第 3 章の分析（3.3.3）は、各国の法制度が共通して、ブラックボックスの因果関係追及の断念と組織責任への読み替え、事業者の自主管理と制裁の組み合わせによる動的変化への対応、そして製品・テスト・組織を統合するセーフハーバーの構築という方向性に向かいつつあることを示した。本節では、これらの知見を踏まえ、UHAI の法制度設計において検討すべき論点を、第 2 章で整理した事前規制・法執行・責任法制の三つの次元に即して整理する。

4.3.1 事前規制：ブラックボックスの法的受容と動的許認可

現行の許認可制度は、申請者がシステムの仕様を明示し、規制当局がその内容を審査しうることを前提とする。UHAI の三つの技術的特性——ブラックボックス

[46] ISO/IEC 42007 (under development), Information technology — Artificial intelligence — Application management system requirements for organizations providing or using AI.

性、動的変化、ゴール設定の困難――は、いずれもこの前提を揺るがすものであり、事前規制の再設計に関して以下の論点が浮上する。

第一に、ブラックボックスの法的受容に関する論点である。4.1.3 で論じた V&V 技術（形式手法によるニューラルネットワーク検証等）が成熟した場合、仕様の明示に代えて安全動作範囲の技術的保証を許認可の要件とすることは法的に受容しうるか。この問いは、許認可制度の前提そのものの再検討を迫るものである。

第二に、動的許認可の可能性に関する論点である。第 3 章で分析した PCCP は、「事前に計画された範囲内の変更」を許容する制度であるが、UHAI の変化は予定も範囲画定もできない場合がある。4.1.3 で論じた実行時モニタリングと許認可制度を連動させ、「安全範囲内の変化は認可を継続し、範囲を逸脱した場合は直ちに運用を停止させる」という動的な許認可の枠組みが法的に成立しうるかは、検討に値する。このような枠組みは、許認可を「一時点の状態審査」から「継続的なプロセス承認」へと転換するものであり、制度整備が必要となる。

第三に、ゴール設定と許認可基準の関係に関する論点である。4.1.1 で論じたリスク許容水準が社会的に合意された場合、その基準を満たすことを許認可の要件とすることは制度的に可能であるが、基準が定性的である限り、許認可の判断には不可避的に裁量が伴う。英国の SoSP（安全原則声明）は、定性的基準をより具体的な安全原則に分解する試みであるが、最終的な技術仕様への翻訳は開発者に委ねられている。この「意味論的ギャップ」をどこまで制度的に埋められるか――そしてどこからは開発者の判断に委ねざるをえないか――の線引きは、事前規制の実効性に直結する論点である。

4.3.2 法執行：自主管理と第三者監視の制度設計

第 2 章で指摘したとおり、UHAI に対する法執行は、ブラックボックス性による監視の困難、動的変化による違反時点の特定困難、そしてゴール設定の困難による執行基準の不在という三重の問題に直面する。第 3 章の分析は、各国がこれらの課題に対し、「外部からの全面的監督」ではなく「事業者の自主管理＋報告義務＋違反時の制裁」の組み合わせによる実効性の確保に向かっていることを示した。

この方向性を UHAI に適用する場合、以下の論点が検討されるべきである。

第一に、自主管理の実効化のための制度的担保である。第 3 章で分析した英国の Duty of Candor（誠実義務）は、事業者に安全上の問題の自主的報告を義務付け、虚偽報告や隠蔽に対してはシニアマネージャー個人に刑事責任を課すことで、企業と個人の間に囚人のジレンマに類似する構造的緊張を創出している。この制度設計は、事業者の自主管理に依拠せざるをえない状況における情報開示の実効性

を確保する一つのモデルであるが、UHAI においては事業者自身がシステムの変化を十分に把握できるとは限らないという追加的な課題がある。

第二に、外部監視の補完的役割である。事業者の自主管理だけでは限界がある場合、独立した第三者による継続的監視の制度化——たとえば航空事故調査委員会に相当する AI 安全専門機関の設置——が必要となるか否かは、法執行体制の設計に直結する論点である。こうした機関には、AI の技術的評価能力、事故の原因調査能力、そして規制当局への勧告権限が求められるが、その設置・運営コストと便益のバランスも考慮される必要がある。

第三に、事後検証のインフラ整備である。自動運転における DSSAD（走行データ記録装置）の義務化は、事後検証のための記録を確保する試みであるが、UHAI においては入出力のログのみで判断過程の再構成がどこまで可能であるかは不確かである。「完全な再構成」を断念した上で、「十分な事後検証」を可能にする記録要件をどう設定するかは、法執行と次節で論じる責任法制の双方に関わる制度設計上の課題である。データ記録義務が存在しても、その開示が適切に行われなければ事後監視モデルは機能不全に陥る。記録義務と開示義務の制度的接続もまた重要な論点である。

4.3.3 責任法制：セーフハーバーの統合的設計と新たな責任類型

UHAI が損害を引き起こした場合の法的責任の在り方は、評価・認証と並んで UHAI の社会実装を左右する重要な制度的条件である。第 2 章で分析したとおり、ブラックボックス性による因果関係の立証困難、動的変化と「引渡し時の欠陥」概念の齟齬、そしてゴール設定の困難に起因する過失判断基盤の不確定性が、責任法制の機能不全を引き起こしうる。第 3 章の分析は、各国がこれらの課題に対し、組織責任への読み替え、ライフサイクル管理の義務化、そして製品・テスト・組織を統合するセーフハーバーの構築という方向性で対処しつつあることを示した。以下では、UHAI の責任法制の設計において検討すべき論点を整理する。

第一に、セーフハーバーの統合的設計に関する論点である。第 3 章で確認したとおり、英国 AVA の due diligence defence（合理的措置の抗弁）は、ASDE が認可要件を充足し誠実義務を遵守し合理的な安全管理措置を講じていた場合に刑事責任を免れうる仕組みであり、「コンプライアンスが防御となる（compliance as defence）」構造を制度化したものである。PCCP における「性能基準＋検証手法＋合格基準」の事前承認、英国の SoSP による C&C 基準の具体化も、セーフハーバーの輪郭を画定する試みと理解できる。これらの知見は、UHAI におけるセーフハーバーが「製品がどのような性能を示すべきか」「それをどのようなテスト環境で検証すべきか」「組織としてどのような管理体制を整備すべきか」という三つの次元を統合する形で設計されるべきことを示唆している。しかし、セーフハーバーの外延が不明確であれば、後知恵バイアス（Hindsight Bias）——事故発生

後に裁判所が事後的に安全基準を再定義するリスク――の問題は依然として残る。セーフハーバーの範囲と限界をいかに明確に画定し、法的予見可能性を確保するかは、責任法制設計の中核的論点である。

第二に、因果関係の立証困難への対応に関する論点である。第 3 章のいずれのケーススタディにおいても、ブラックボックス内部の因果の鎖を直接的に解明する方法は提示されていない。英国 ASDE の法的擬制は、因果関係の追及そのものを回避し結果を組織に帰属させるという回答を示したが、これは擬制の前提となる組織的責任主体の存在に依拠しており、責任主体の設計と不可分である。ここで検討すべきは、因果関係の推定や立証責任の転換といった手続的対応にとどまるのか、それとも因果関係の立証を要件としない責任類型――厳格責任や無過失責任――への移行を検討すべきかという選択である。[47]製造物責任の枠組みが UHAI の因果関係問題にどこまで対応しうるかは、今後の運用を注視すべき論点である。

第三に、「引渡し時の欠陥」概念の再構成に関する論点である。製造物責任法における「引渡し時の欠陥」要件は、動的に変化する UHAI との間に本質的な齟齬を生じさせる。引渡し後の学習や環境変化により性能が劣化し、あるいは予期しない判断傾向を獲得した場合に、「いつの時点の状態を基準に欠陥を判断するか」について現行法は明確な回答を持たない。英国 ASDE モデルにおける継続的安全性義務は、責任の基準時を「引渡し時」からライフサイクル全体へと拡張する方向性を示唆している。この方向に進む場合、開発者・提供者・運用者の間の責任分担をライフサイクルの各段階に応じてどう設計するかが、サプライチェーン全体にわたる責任法制の再構成を伴う大きな論点となる。

4.4 まとめ：UHAI の恩恵を最大限に享受できる社会に向けて

以上は、いずれも今後議論されるべき論点を整理したものであって、確定的な回答を示したものではない。しかし、本報告書の分析を通じて、評価・認証・法制度の三層のそれぞれについて、制度構築の方向性は見えつつある。以下、本報告書の総括として、これを簡潔に整理する。

第一に、評価については、UHAI の三つの技術的特性に対応する新たな評価パラダイムを構築する、という方向性である。ブラックボックス性に対しては、AI の内

[47] AI に対する厳格責任の導入可能性に関する議論については、Vladeck, D.C. (2014). Machines Without Principals: Liability Rules and Artificial Intelligence. Washington Law Review, 89(1), 117‑150 参照。

また、EU における AI 責任法制の議論の概観として、European Parliament (2020). Civil Liability Regime for Artificial Intelligence: European Added Value Assessment, PE 654.178 参照。

部を解明するのではなく、形式手法やレッドチーミング等の新たな V&V 手法を活用して AI の「振る舞い」を外側から検証する技術を標準化する。動的変化に対しては、出荷前の検証と市販後のリアルワールドパフォーマンス監視を一体的に設計し、ライフサイクル全体を通じた継続的評価の仕組みを整備する。ゴール設定の困難に対しては、「絶対的な安全」を求めるのではなく、従来の人間によるオペレーションとの相対的かつ統計的な比較を軸に、多様なステークホルダーが参加する透明なプロセスでリスク許容水準を策定し、その合意を評価基準の出発点とする。

第二に、認証については、上記の評価基準を実効的に担保するための認証制度を構築する、という方向性である。ブラックボックス性に対しては、製品の入出力検証だけでなく、それを開発・運用する組織のガバナンス体制を併せて審査する「ジョイント認証」の枠組みを構築する。動的変化に対しては、変更のたびに一から審査をやり直すのではなく、変更管理のプロセスそのものを事前に承認するアプローチを発展させる。ゴール設定の困難に対しては、社会的に合意された基準を認証の判定基準に接続し、認証機関がそれを独立に検証しうる能力を構築する。

第三に、法制度については、評価と認証の成果を法的に受容し、事業者の予見可能性を確保する、という方向性である。ブラックボックス性に対しては、内部構造の説明に代えて安全動作範囲の技術的評価を許認可の要件として受容し、適切な組織体制の維持をセーフハーバーとして法的に位置づける。動的変化に対しては、事業者に継続的な安全管理と自主報告を義務付けつつ、情報提供への適切なインセンティブ付けを行うことによってその実効性を確保する。ゴール設定の困難に対しては、社会的に合意された基準を達成していることを事業者が証明できるようなテスト環境や認証制度を整備する。

これら三つの方向性に共通するのは、評価・認証・法制度の三層を個別に設計するのではなく、三層が一体として機能するよう統合的にデザインしなければならない、ということである。評価基準の合意が認証の基盤となり、認証の取得がセーフハーバーの根拠となり、セーフハーバーの明確化が事業者の予見可能性を高めてイノベーションを促進する――この好循環を実現することが、第 2 章で指摘した「悪循環」を断ち切る鍵である。

UHAI は、人間の能力の限界を超え、社会により大きな安全と便益をもたらす可能性を秘めている。その可能性を現実にするために、私たちは「人間が監視すれば安全」という従来の前提を超え、技術の進化に適応しうる新たな制度の枠組みを構想しなければならない。本報告書で整理した論点が、その議論の出発点となることを期待する。

＜表：UHAI の特徴（横軸）及びそれに対応する制度設計の方向性（縦軸）＞

	ブラックボックス性	動的変化	ゴール設定の困難性
評価	AI の内部解明を断念し、形式手法・レッドチーミング等の新たな V&V 手法で AI の「**振る舞い」を外側から検証**する技術を標準化	出荷前検証と市販後の**リアルワールドパフォーマンス（RWP）監視を一体設計**し、ライフサイクル全体を通じた継続的評価を整備	「絶対的な安全」ではなく、人間によるオペレーションとの**相対的・統計的比較**を軸に、多様なステークホルダー参加の透明なプロセスでリスク許容水準を合意
認証	製品の入出力検証＋組織のガバナンス体制を**セットで審査する「ジョイント認証」**の枠組みを構築	変更のたびに再審査するのではなく、**変更管理プロセスそのものを事前承認**するアプローチを発展	社会的に合意された基準を**認証判定基準に接続**し、認証機関がそれを独立に検証しうる能力を構築
法制度	内部構造の説明に代え、**安全動作範囲の技術的保証を許認可要件**として受容。適切な組織体制の維持を**セーフハーバー**として法的に位置づけ	事業者に**継続的安全管理と自主報告を義務**付けつつ、情報提供への適切なインセンティブ付けにより実効性を確保	合意された基準の達成を事業者が証明できる**テスト環境・認証制度を整備**し、法的予見可能性を確保するセーフハーバーを構築

以上

2025年度 AIの活用における課題と施策に関する研究会

【 構 成 員 名 簿 】（敬称略・五十音順）

※報告書執筆者

■ 主査

- **羽深 宏樹***　国立大学法人京都大学大学院法学研究科
  附属法政策共同研究センター　特任教授
  弁護士

■ 委員

- **伊藤 公一**　PwC Japan有限責任監査法人　パートナー
  AI監査研究所所長　公認会計士
- **稲谷 龍彦**　国立大学法人京都大学大学院法学研究科　教授
- **落合 孝文**　渥美坂井法律事務所 プロトタイプ政策研究所 所長
  シニアパートナー弁護士
- **杉村 領一**　国立研究開発法人産業技術総合研究所
  情報・人間工学領域　連携推進室
  チーフ連携オフィサー
- **髙橋 久実子**　株式会社三菱総合研究所
  社会インフラ事業本部　都市インフラDXグループ
  主任研究員
- **陳 冠瑋***　国立大学法人京都大学大学院法学研究科　特定助教
- **鄭 育昌***　富士通株式会社 データ＆セキュリティ研究所
  シニアリサーチマネージャー
- **富安 啓輔**　株式会社AIメディカルサービス, CTO
- **広瀬 貴之**　国立大学法人京都大学大学院法学研究科 特定講師

■ オブザーバー

- **髙村 博紀**　独立行政法人情報処理推進機構（IPA）
  デジタル基盤センター　デジタルエンジニアリング部
  AIシステムグループ　主任研究員
- David Socol de la Osa David Uriel
  国立大学法人一橋大学社会科学高等研究院 准教授

■ **事務局** 一般財団法人 国際経済連携推進センター（CFIEC）

- **加藤 幹之**　デジタル社会研究所　所長
- **松沢 栄一**　デジタル社会研究所　研究主幹
- **片桐 守雅**　デジタル社会研究所　主任研究員