NTEN-2024-1002

NRA 技術ノート

NRA Technical Note Series

原子力分野における人工知能の動向に係る調査報告

Review of the current status of the development of artificial intelligence and their application in the nuclear field

宮﨑　利行　　野口　法秀
MIYAZAKI Toshiyuki and NOGUCHI Norihide
システム安全研究部門
Division of Research for Reactor System Safety

太田　良巳　　東　喜三郎
OTA Yoshimi　and　AZUMA Kisaburo
地震・津波研究部門
Division of Research for Earthquake and Tsunami

梁田　勇太
YANADA Yuta
シビアアクシデント研究部門
Division of Research for Severe Accident

原子力規制庁
長官官房技術基盤グループ
Regulatory Standard and Research Department,
Secretariat of Nuclear Regulation Authority (S/NRA/R)

令和 6 年 9 月
September 2024

本報告は、原子力規制庁長官官房技術基盤グループが行った安全研究等の成果をまとめたものです。原子力規制委員会は、これらの成果が広く利用されることを期待し適時に公表することとしています。

なお、本報告の内容を規制基準、評価ガイド等として審査や検査に活用する場合には、別途原子力規制委員会の判断が行われることとなります。

本報告の内容に関するご質問は、下記にお問い合わせください。

原子力規制庁　長官官房　技術基盤グループ　システム安全研究部門
〒106-8450　東京都港区六本木 1-9-9 六本木ファーストビル
電　　話：03-5114-2223
ファックス：03-5114-2233

NTEN-2024-1002

原子力分野における人工知能の動向に係る調査報告

原子力規制庁 長官官房技術基盤グループ
システム安全研究部門
宮﨑　利行、野口　法秀

地震・津波研究部門
太田 良巳、東 喜三郎

シビアアクシデント研究部門
梁田 勇太

要　　旨

人工知能（Artificial Intelligence: AI）は近年、大きく能力を高めている。AI の能力を生かし柔軟で効率的な産業活動を目指す取組みが世界的に検討されており、我が国の原子力分野にも AI の適用が見込まれることから、技術基盤グループ内でチームを立ち上げて調査を行った。

IAEA や OECD/NEA では原子力分野での AI 利用を目的として国際協力を進めており、IAEA はその成果として幅広い分野での適用を検討した報告書を発行している。OECD 全体での AI 利用に関する政策的な取組みと併行して、OECD/NEA においては AI の利用検討とともに実践的なベンチマーク解析の機会を加盟国に提供しており、この点は OECD/NEA の活動の大きな特徴である。

EU では包括的に AI を規制する AI 法を 2024 年に成立させ、2026 年には全面的に施行される予定であり、重大なリスクをもたらしうる原子力分野での一部の AI 利用はハイリスク AI として規制の対象となる見通しである。英米は EU とは対照的に、国として AI の法規制を行わず、個々の規制当局が非法的規制を行う方針をホワイトペーパー、大統領令などで定めており、英 ONR や米 NRC 等はそれに従って AI 対応を進めている。

原子力分野での AI 利用に関しては、事業者や大学、研究機関等で原子力施設の監視・運用、設計の最適化、リスク評価、放射線防護・核セキュリティ、材料・構造分野等での検討が進められている。日本国内では、実際に AI を活用した安全保護具の装備確認などで事業者による運用例が報告されているが、リスクを判断するような利用例は見られない。

幅広い分野で AI の利用への期待が示されているものの、AI に特有のリスクへの懸念も示されている。AI のリスクに対処するための仕組みとして AI の利用に指針を与える「AI ガバナンス」が世界全体から個々の企業レベルまで導入されつつある。

NTEN-2024-1002

Review of the current status of the development of artificial intelligence and their application in the nuclear field

MIYAZAKI Toshiyuki, NOGUCHI Norihide
Division of Research for Reactor System Safety

OTA Yoshimi, AZUMA Kisaburo
Division of Research for Earthquake and Tsunami

YANADA, Yuta
Division of Research for Severe Accident
Regulatory Standard and Research Department,

Secretariat of Nuclear Regulation Authority (S/NRA/R)

Abstract

Capability of Artificial intelligence (AI) has been greatly developed in recent years. Efforts aiming at flexible and efficient industrial activities by utilizing AI are being considered worldwide, As AI is expected to be used in the Japanese nuclear field, a survey team was formed within the Regulatory Standard and Research Department to collect and review related information.

IAEA and OECD/NEA are promoting international cooperations for utilization of AI in the nuclear field, and the IAEA has issued a report which examines application of AI in various fields as a result of such cooperations. In parallel with implementation of OECD's policy initiatives, OECD/NEA provides its member countries with opportunities to examine AI utilizations and perform practical benchmark analysis, which is a major feature of the OECD/NEA's activities.

In the EU, AI Act that comprehensively regulates AI was passed in 2024 and is scheduled to be fully enforced in 2026. Some AI applications in the nuclear field, which could pose significant risks, are expected to be subject to regulation as high-risk AI. In contrast to the EU, the U.K. and the U.S. have established their policies in white papers, presidential directives, and other documents stating that individual regulatory authorities should conduct non-legal regulation of AI.

In the field of nuclear energy, operators, universities, research institutes, and other organizations are studying the use of AI for monitoring and operation of nuclear facilities, design optimization, risk assessment, radiation protection and nuclear security, materials and structures, and other areas. In Japan, there have been reports of actual applications of AI by operators to check

the equipment of safety protection equipment, but there have been no examples of AI applications to determine risk.

While expectations are expressed for utilization of AI in various fields, concerns are also expressed about the risks inherent in AI. “AI governance” is being introduced from the global level to the individual company level to provide guidance for the use of AI as a mechanism for addressing AI risks.

目　次

表　目　次

図　目　次

略　語　表

AGI	Artificial General Intelligence（汎用型人工知能、強い AI）
AI	Artificial Intelligence（人工知能）
AIRMF	AI Risk Management Framework（AI リスクマネジメントフレームワーク）
AISC	AI Steering Committee（AI 運営委員会）
AISI	AI Safety Institute（AI セーフティー・インスティテュート）
ANI	Artificial Narrow Intelligence（特化型人工知能、弱い AI）
BERT	Bidirectional Encoder Representations from Transformers（Google による大規模言語モデル）
CAP	Corrective Action Program（改善処置活動）
CE	Conformité Européenne（ヨーロッパの適合性）
CEIMIA	Centre d'expertise international de Montréal en intelligence artificielle
CFO	Chief Financial Officer（最高財務責任者）
CNN	Convolutional Neural Network（畳み込みニューラルネットワーク）
CNSC	Canadian Nuclear Safety Commission（カナダ原子力安全委員会）
CR	Condition Report（原子力安全に影響を及ぼすおそれのある情報）
CRS	Congressional Research Service（米国議会調査局）
DHS	United States Department of Homeland Security（米国国土安全保障省）
DOE	United States Department of Energy（米国エネルギー省）
EC	European Commission（欧州委員会）
EO	Executive Order（行政命令（米国））
EPRI	Electric Power Research Institute
EU	European Union（欧州連合）
FLOPS	FLoating-point Operations Per Second
FLOPs	FLoating-point OPerations
GIS	Geographic Information System（地理情報システム）
GPT	Generative Pre-trained Transformer（OpenAI による大規模言語モデル）
GPU	Graphics Processing Unit（グラフィックスプロセッサ）
IAEA	International Atomic Energy Agency（国際原子力機関）
ICBMs	Independent Confidence Building Measures（独立した信頼性向上対策（ONR））
ILSVRC	The ImageNet Large Scale Visual Recognition Challenge（ImageNet が 2010 年から 2017 年まで開催した画像認識技術コンテスト）
INL	Idaho National Laboratory（アイダホ国立研究所）

IPA	Information-technology Promotion Agency（独立行政法人情報処理推進機構）
ITU	International Telecommunication Union（国際電気通信連合）
LSTM	Long Short-Term Memory（長・短期記憶、RNN の一種）
ML	Machine Learning（機械学習）
MOU	Memorandum of Understanding（了解覚書）
NDE	Nondestructive Evaluation（非破壊評価）
NIST	National Institute of Standards and Technology（米国立標準技術研究所）
NLP	Natural Language Processing（自然言語処理）
NPP	Nuclear Power Plant（原子力発電所）
NRC	U. S. Nuclear Regulatory Commission（米国原子力規制委員会）
OECD	Organisation for Economic Co-operation and Development（経済協力開発機構）
OECD/NEA	OECD/Nuclear Energy Agency（経済協力開発機構/原子力機関）
OMB	Office of Management and Budget（米国行政管理予算局）
ONR	Office for Nuclear Regulation（英国原子力規制局）
OPM	United States Office of Personnel Management（米連邦政府人事管理局）
OSTP	Office of Science and Technology Policy（米国科学技術政策局）
PRA	Probabilistic Risk Assessment（確率論的リスク評価）
RC	Reinforced-Concrete（鉄筋コンクリート）
RCA	Root Cause Analysis（根本原因分析）
PE	Production Excellence（製造の卓越性（ONR））
RIC	Regulatory Information Conference（規制情報会議（NRC））
RNN	Recurrent Neural Network（回帰型ニューラルネットワーク、内部記憶を持ち、時系列データに適したニューラルネットワーク）
SAPs	Safety Assessment Principles（原子力安全規則（ONR））
SRMA	Sector Risk Management Agency（セクター別リスク管理機関（米国））
TFEU	Treaty on the Functioning of the European Union（EU の機能に関する条約）
XAI	eXplainable AI（説明可能な AI）

1. はじめに

人工知能（以下「AI」という。）を適用した柔軟で効率的な産業活動を目指す取組みが世界的に行われており、原子力分野への適用についても特に欧米を中心に各国で検討が進められている。

今般、第 39 回原子力規制委員会（令和 5 年 10 月 25 日）において原子力規制委員会委員より AI（及び先進製造技術（注 1）（Advanced Manufacturing Technology））に関する各国の規制機関、国内外の産業界及び国際機関の状況について調査依頼があったことを受け、国際機関（国際原子力機関（以下「IAEA」という。）、経済協力開発機構／原子力機関（以下「OECD/NEA」という。）等）並びに欧米等の規制機関及び産業界における AI に関する動向について、技術基盤グループ内でチームを立ち上げて調査を行い、関連情報を収集・整理した。

本技術ノートでは、第 2 章において AI についての基礎事項を、第 3 章において原子力分野での国際機関の AI への対応、欧州連合（EU）での予想される規制方針、英国原子力規制局（以下「ONR」という。）や米国原子力規制委員会（以下「NRC」という。）の対応の概略を、第 4 章において原子力や関連する分野での AI 利用についての調査結果を示す。また AI の利用に際しては様々なリスクが想定されるが、そのリスクの概要を第 5 章に、国際機関、英国、米国、EU 及び日本が、AI のリスクに対処するための仕組みとして制度を整備している AI ガバナンスの概要を第 6 章に示す。AI に関する参考文献は、執筆中に新たに発表されたものもできるだけ含めるよう努めたが、全てを網羅するものではない。これは、AI 利用やその検討が急速に進められ、あわせて各国政府が AI ガバナンス体制を構築しているという状況にあるためである。本文同様、日々更新されていくものであることに留意されたい。

2. AI について

2.1 AI の定義

2019 年 3 月に内閣府の内閣府統合イノベーション戦略推進会議が定めた「人間中心の AI 社会原則」[2-1] では「（AI の）定義については研究者によっても様々な考え方があり、現在のところ明確な定義はない」と述べている。2024 年 4 月の総務省と経済産業省が策定した「AI 事業者ガイドライン」[2-2] でもその立場を受け継いでおり、日本政府として公式に AI を定義することは避けている。英国なども同様の立場を取っているが、一般的には OECD の「人工知能に関する理事会勧告」[2-3] の

（注 1） 先進製造技術に関する調査結果は、NRA 技術ノート NTEN-2024-1001「先進製造技術の開発及び原子力分野への適用の現状に関する調査」として報告する。

AI システムとは、人間が定義した一定の目的のために、実環境あるいは仮想環境に影響を及ぼす予測、推薦又は意思決定を行う機械ベースのシステムである。AI システムは様々なレベルの自律性を備えて稼働するよう設計されている。

という定義（日本語訳は総務省による非公式翻訳[2-4]による）が代表的な定義だと受け止められているようであり、本報告でもこの定義に基づき情報を整理した。

なお、AI の定義について様々な考え方があるというのは、AI のひとつの領域である知能（Intelligence）の定義について様々な考え方があるためである。AI の基礎を作った Turing は「人工知能（AI）」という言葉こそ使用していないものの、早い段階から計算機による知性の実現可能性を検討していた[2-5]。彼は、チューリングマシーン[2-6]に代表される論理的計算機（logical computing machines: LCMs）や当時現実のものとなりつつあった実用計算機（practical computing machines: PCMs）とは別に教育、あるいは快楽と苦痛を通じて成長していく機械学習（注 2）的な機械についても考察している。また 1950 年の論文[2-7]で、Turing は「機械は考えることができるのか？」という疑問とともに、機械が人間的であるかどうかを判定する「チューリングテスト」を提案しており、その考えに従うなら AI は「人間的」でなければならない。一方、1997 年にチェスで世界チャンピオンのカスパロフを破った IBM の Deep Blue や、2017 年に囲碁で世界トップ棋士の柯潔を破った Google DeepMind の AlphaGo は、専門分野では人間よりも高い能力を持つが、それ以外のタスクでは人間的とは言えない。そこでより厳密に、特定のタスクでのみ力を発揮する AI を特化型人工知能（Artificial Narrow Intelligence: 以下「ANI」という。）あるいは「弱い AI」、人間が実現可能なあらゆるタスクをこなせる AI を汎用型人工知能（Artificial General Intelligence: 以下「AGI」という。）、あるいは「強い AI」などと区別することがある[2-8]。2024 年現在では AGI と呼べるような AI は存在せず、最新の生成 AI が AGI への一歩であるのかどうかが議論になっている[2-9]。

本報告の「AI」は特別に断らない限り、「ANI」のことである。

2.2 手法による分類

本節では AI の手法について説明する。

2.2.1 探索による手法

この手法は、初期状態から起こり得る状態変化を並べて、望ましい状態の並びを選んでいく手法である（図 2.1）。図の例では「A」というノードを始点として順次探索を深めていっている。途中、「E」や「G」のように赤の×印で表されているノードは見込みが無いのでそこで探索を打ち切る。この手法は、コンピューターにチェスをプレイさせる手法として、Shannon[2-10]や Turing[2-11]によって AI という言葉が生まれる以前から提案されている。

（注 2） 2.2.3 参照。

単純な問題（ゲーム）であれば全ての場合を網羅して最善の選択肢を選ぶこと（つまり、その問題を完全に解析すること）が可能であるが、ある程度以上の規模のゲームでは現実的には不可能である。Shannon は、チェスの場合では10^{120}通りの組み合わせがあると概算し（あくまでも概算なので、後年の見積とは若干異なる）、全ての場合を網羅することは不可能であることから、駒の価値などを基にした評価関数により指し手を選ぶ手法を提案している。この手法で最も注目すべき成果を収めたのは IBM のチェスコンピューターDeep Blue で、松原 [2-12] によると、512 台のチェス専用チップというハードウェア的な対応もさることながら、序盤と終盤の徹底的なデータベース化や反復深化により、Shannon の提案した探索手法を「ときに 50 手 60 手先まで読む」までに高度化して 1997 年にチェスの世界チャンピオンを破った。

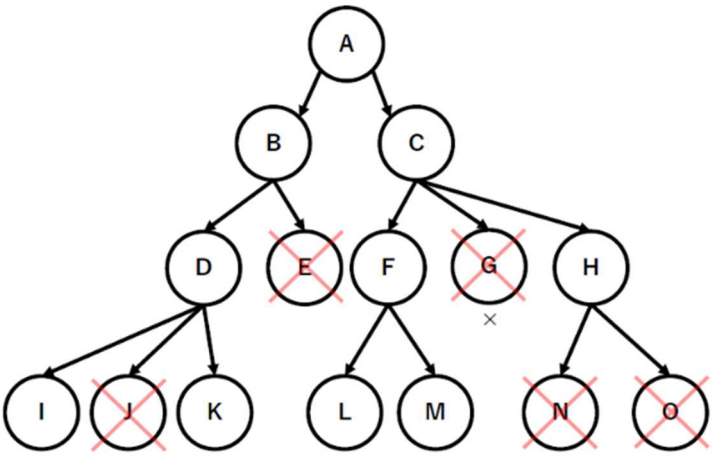


図 2.1　探索による手法の例

Fig. 2.1　An example of tree search methods

2.2.2　エキスパートシステム

チェスや将棋といったゲームは、ルールに沿った行動しか取れないので、探索による手法である程度の AI を作ることができるが、その手法で現実の問題を解くのは難しい。そこで専門家の知識を用いて推論を行い、現実の問題を解決するシステムとして考案されたのがエキスパートシステムである。上野 [2-13] は、エキスパートシステムを「問題領域の専門家（エキスパート）から獲得された専門知識を用いて推論を行い、専門的に高度な現実の問題を、専門家と同等のレベルで解決する知識システムをいう」と定義し、構成要素として

- 専門知識を表現しそれを統合的に管理する機構である知識ベース
- 知識ベース内の知識を利用して、推論を実行するための推論機構
- ユーザとの応答をスムーズに行うためのユーザ・インタフェース
- エキスパートから専門知識を獲得し、知識ベースを構築する作業を支援するための知

識獲得支援機構

- ユーザの要求に応じて、推論で導いた結論の根拠を説明するための推論過程説明機構など

を挙げている。図 2.2 に簡単なエキスパートシステムの例を示す。この例では質問を重ねることで、データベースから人物の特定を行っている。

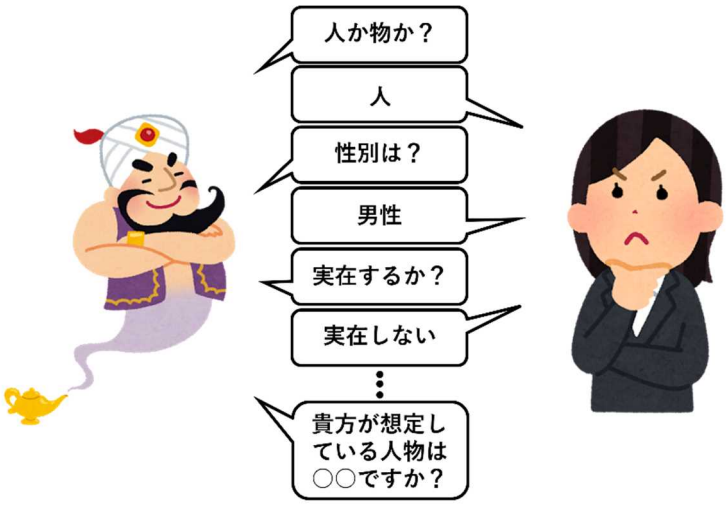


図 2.2　エキスパートシステムの例

Fig. 2.2　An example of expert systems

上野は、最初のエキスパートシステムとして 1965 年にスタンフォード大学で研究が開始された、化学構造式同定システム DENDRAL [2-14] を挙げている。その後 1970 年代に米国の大学で MYCIN（スタンフォード大学、感染症の診断と投薬 [2-15]）、CASNET（緑内障の診断と治療 [2-16]）、INTERNIST（内科診断 [2-17]）、HEARSAY-Ⅱ（カーネギーメロン大学、音声認識 [2-18]）などのシステムが開発された。

日本でも 1982 年に旧通商産業省所管の新世代コンピュータ技術開発機構による「第五世代コンピューター開発プロジェクト」[2-19] が契機となって、AI に対する産業界の関心が高まり、多くの企業が導入したり、導入を検討したりした [2-20]。原子力分野も例外ではなく、設計支援ツール等の検討例が報告されている [2-21, 2-22, 2-23, 2-24]。

松尾 [2-25] によるとエキスパートシステムの課題は、

- 専門家からヒアリングして知識を取り出すためのコスト
- 知識の数が増えると、お互いに矛盾していたり、一貫していなかったりするので、適切に維持管理する必要がある
- 高度な専門知識が必要な限られた分野ではともかく、「常識レベルの知識」を記述するのが難しい

などである。また、寺野 [2-26] は「プラントの故障診断を対象とした研究開発例は非常に多い。ところが診断システムで実用化されたものは非常に少ない。」とし、その理由として「特

に工学的なシステムを対象とする場合、経験則で判断できるような故障は設計段階で排除されるのが普通であり、設計当初予想できなかったような問題が故障として発生するという事情による。」と説明している。

2.2.3　機械学習（シャローラーニング）

エキスパートシステムでは専門家の知識を人間が整理してコンピューターに与えていたが、データをコンピューター自身に学習させるのが機械学習（Machine Learning）である。Samuel が IBM701 のために作成したチェッカープログラム [2-27] は最初期の機械学習プログラムだと考えられている。このプログラムは、探索による手法を主としながらも、プレイ中に遭遇した局面を探索結果とともに記憶しているので、一度遭遇した局面では通常よりも先の局面まで探索することができる。Samuel のプログラムでは評価関数のパラメーター修正も行っていたようである。

「機械学習」と呼ばれる手法の範囲には様々な見解があるが、付録 1 で例を示す。後述の深層学習（ディープラーニング）はニューラルネットワーク（注 3）の層を深くした特別な場合なので機械学習の一種ではあるが、機械学習とは分けて考えられることが多い。その区別を明らかにするために深層学習ではない機械学習を「シャローラーニング」と呼ぶこともある。ニューラルネットワークも含む機械学習の詳細については、Raschka ら [2-28] などに書かれている。

2.2.4　深層学習（ディープラーニング）

層を深くすればニューラルネットワークの表現力が向上し、予測性能も向上することが予想されていたが、深層学習と呼ばれる技術が広く使われるようになったのは 2010 年以降である。Chollet [2-29] によると 2010 年代以降の深層学習の流行の要因は

① ハードウェア

深層学習で大量に実行する積和計算を並列で高速に実行できる、Graphics Processing Unit（GPU）と、NVIDIA 社製の GPU 用のプログラミングインターフェースの Compute Unified Device Architecture（CUDA）により、これまでは学習が困難だったサイズのモデルも学習できるようになった。

② データ

インターネットの台頭によって機械学習用の大規模なデータセットの収集と配布が可能になった。特に重要なのは 1000 種類の画像カテゴリでタグ付けされた 140 万枚の画像で構成された ImageNet [2-30] で、毎年開催している ILSVRC（ImageNet Large Scale Visual Recognition Challenge）[2-31] というコンテストが深層学習の発展に重要な役割を果たしている。

（注 3）　付録 1 を参照。

③ アルゴリズム

層数の深いニューラルネットワークは多くの情報を記憶することができるが、勾配消失と呼ばれる学習が進まなくなる問題などにより使用が難しかった。しかし、活性化関数の改善、重みの初期化方法の改善、最適化手法の改善などのアルゴリズムの改善により、10 層以上のモデルの学習が可能になった。さらに、バッチ正規化、残差接続等の高度な手法の発見により非常に深いモデルでも学習ができるようになった。

④ 投資

AI、特に深層学習への投資が爆発的に増え、研究のペースが上がった。

⑤ 大衆化

Python 上で深層学習の開発を行うことのできるライブラリ（Theano、Tensor-Flow、Keras など）が開発された。

としている。

特に画像認識の分野ではパターン認識に長けた畳み込みニューラルネットワーク（Convolutional neural network: 以下「CNN」という。）と呼ばれる技術により飛躍的に性能が向上した。ただし、CNN 自体は 1980 年頃にネオコグニトロン（Neocognitron）として福島が提案している [2-32, 2-33]。世界に深層学習と CNN の威力を知らしめたのは 2012 年の ILSVRC [2-31] で優勝した AlexNet [2-34] で、他のモデルのエラー率が 26 %以上のところ、15 %台というそれまでの記録を 10 %以上更新する画期的な性能を示した。AlexNet は、全パラメーター数が 6000 万という深いニューラルネットワークで、ディープニューラルネットワークが普及する契機となり、またそれ以降は画像認識・分類モデルの改良が深層学習の開発を牽引した。

一方、自然言語処理（natural language processing: NLP）（参考文献として例えば斎藤 [2-35]）は機械学習の伝統的な研究分野だったが、リカレントニューラルネットワーク（recurrent neural network: RNN）と呼ばれるニューラルネットワークが 2010 年代半ばから用いられるようになり、深層学習の利用が一般的になった。この分野では 2017 年に Google の研究者ら [2-36] が発表した Transformer が、注意機構（attention mechanism）という仕組みを中心としたモデルにより、時系列データの逐次処理を不要とし、並列化を容易とした。Transformer により、より大きなデータセットでの学習が可能となり、Bidirectional Encoder Representations from Transformers (BERT) [2-37] や Generative Pre-trained Transformer (GPT) [2-38] といった大規模言語モデルが開発されている。現在ではこれらの大規模言語モデルが深層学習開発の中心となっている。2020 年に OpenAI が公開した GPT-3 はモデルのパラメーター数が 1750 億であり [2-39]、タスクは異なるが AlexNet と比較して飛躍的に巨大なモデルとなっている。

大規模言語モデルでは学習データの次元を削減して潜在ベクトルと呼ばれるベクトルに変換する。この変換によって入力データ全体を表現し直したものを潜在空間と呼ぶ。大

規模言語モデルに何か文字列を与えると、入力された文字列に関係のある潜在ベクトルを潜在空間から取り出し、学習データの空間に逆変換して文字列として出力する。

同様に画像、音楽などの潜在空間を学習すると、与えた入力（文字列など）に従って画像、音楽を生成する。このようなタスクを行う深層学習を生成型 AI（又は生成 AI）と呼ぶが、これらのモデルは、学習した潜在空間からサンプルを抽出するので、完全に新奇な画像や音楽を生成するわけではない（参考文献として例えば斎藤 [2-40]、アンドリューら [2-41]）。

2.3 機械学習の学習方法による分類

機械学習は学習方法によって、教師あり学習、教師なし学習、強化学習の 3 種類に分類されることが多い。ここではそれぞれについて解説する。

2.3.1 教師あり学習

画像やデータ、音声などのデータに、対応するラベルを付けた学習データ（訓練データ）でモデルを訓練し、未知のデータや将来のデータのラベルを予測できるようにするタスクである。ラベルが離散値の場合は分類、連続値の場合には回帰とも呼ばれ、多くの機械学習モデルはどちらのタスクにも対応できる。分類で最も単純なのは陽性・陰性に分類する二値分類だが、多くのクラスに分類する多値分類も良く使われる。

2.3.2 教師なし学習

教師なし学習ではラベル付けされていないデータや構造が不明なデータを扱う。よく使われるタスクとしては、クラスタリング、次元削減が挙げられる。このうち、クラスタリングではデータを、その特徴からいくつかのグループに分類する。クラスタリングの古典的なアルゴリズムとしては k-means（k 平均法）が有名である。次元削減では高次元のデータをより低い次元に圧縮する。古典的には主成分分析（principal component analysis: PCA）が有名である。

k-means も主成分分析も、機械学習モデルとは独立に成立・発展してきた手法であるが、機械学習モデルと組み合わせて使用されることも多い。また、現在では深層学習によるクラスタリングや次元削減も行われている。

k-means 及び主成分分析については Raschka ら [2-28] などを参照。

2.3.3 強化学習

強化学習では環境とのやり取りを通じて性能を改善するシステム（エージェント）を開発する。ゲームをプレイするAI、自動車を運転するAIなどが含まれ、Samuelのチェッカープログラム [2-27] が最も初期の例である。強化学習では自ら環境に対して作用を行いながら、その行動の良し悪し又は状況の価値などを学習していく（参考文献として斎藤 [2-42]）。

強化学習でも従来は機械学習的な手法が用いられてきたが、現在では深層学習によって

性能が改善されている（深層強化学習）。DeepMind 社が開発した AlphaGo [2-43] は深層学習を用いて囲碁をプレイする AI であるが、従来の機械学習では困難であると考えられていた人間のチャンピオンに対する勝利を実現している。

2.3.4 半教師あり学習

以上の 3 つの分類は広く使われているものであるが、現在では必ずしもその枠に収まらないものも多い。例えばデータ数が非常に多い場合には「半教師あり学習」と呼ばれる手法が用いられる。これは、例えば少数のラベル付きデータで学習した分類器をラベルなしのデータに適用し、そこで確度が高いと考えらえるデータと分類を学習データとしてモデルをさらに訓練するような手法である。大規模言語モデルでも教師なし学習と教師あり学習の両方の手法を用いた「半教師あり学習」で学習することが多いようである。

また、前述の AlphaGo は最初の段階で人間の対局した棋譜から次の一手を学習するという「教師あり学習」を行っているのに対し、その改良版である AlphaGo Zero [2-44] で、囲碁のルールだけを教えたプログラムが自己対局だけで学習を行うという「教師なし学習」で、数日で AlphaGo よりも高い性能に到達した。ここで注目すべきは「教師あり学習」の場合人間がラベル付けをしたデータなどから学習を行うため、AI も人間の偏見、先入観を受け継いでしまうのに対し、「教師なし学習」では人間の先入観などに縛られないため、最終的にはより公平で高性能な AI を開発できる可能性があることである。

3. 国際機関や各国規制機関の取組み

本章では国際連合、IAEA、OECD/NEA といった国際機関における原子力分野での AI 対応と、各国原子力規制当局の AI 対応について述べる。これまでに確認した中では、各国原子力規制当局については、英国 ONR と米国 NRC が AI 対応について、まとまった文書を発表している。EU 加盟国の規制当局の方針等は発表されていないが、原子力分野も水平的に適用される EU の AI 法の適用対象となるため、同 AI 法に沿った原子力規制を行うものと考えられる。そこで本報告では EU の AI 法から予想される EU 加盟国の原子力規制当局の対応と、これまでの ONR、NRC の公表資料等から予想される今後の対応について述べる。

3.1 国際機関

3.1.1 国際連合

2015 年の国際連合（以下「国連」という。）総会で、人間、地球及び繁栄のための行動計画として、「持続可能な開発のための 2030 アジェンダ」[3-1, 3-2] を採択した。本アジェンダには、2030 年までに達成するべき持続可能な開発目標（SDGs）として、17 の目標が示されている。本アジェンダには、すべての国及びすべてのステークホルダーが、協同的なパートナーシップの下でこの行動計画を実行することを宣言している。

国連を構成する各機関では、17 の目標の達成に有効な手段となりうる技術として、AI を利用することを検討している。2019 年には、国際電気通信連合（ITU）の主導で、AI の課題対処能力向上のための戦略的アプローチとロードマップ [3-3] が作成されている。このような各国連機関が有する AI に関する専門的な知見を集約するため、国連の最高責任者調整会議とその上位委員会は、AI ワーキンググループ（以下「IAWG-AI」という。）を発足させた。IAWG-AI の活動として、国際連合教育科学文化機関（UNESCO）と情報通信技術事務局（OICT）は、国連機関における AI の倫理的な利用に係る原則 [3-4] を作成した。

ITU は、2017 年から AI に関する国際サミット「AI for Good」[3-5] を開催している。「AI for Good」には 40 を超える国連機関が参画し、各機関の活動の概要がレポート [3-6] として報告されている。「AI for Good」には IAEA も参画しており、原子力分野での AI 利用を検討するためのプロジェクトを進めている。

3.1.2 国際原子力機関（IAEA）

原子力分野における AI の利用は、原子力科学技術、放射線防護、核セキュリティ及び保障措置等、多岐に渡るため、IAEA は様々なプロジェクトを「AI for Good」の枠組みで進めている。IAEA は、これらの分野における将来的な活動の優先順位と、IAEA として取り組むべき内容を特定するため、2021 年に専門家会合を開催し、レポート [3-7] を発行している。

同レポートでは、AI や機械学習の手法が、平和、健康、繁栄に貢献するという IAEA の目標に向けて、原子力利用、科学、技術の分野を加速することができるとしており、恩恵

を受ける分野として、人類の健康、食料及び農業、水及び環境、原子力科学及び核融合研究、原子力発電、核セキュリティと放射線防護、保障措置があるとしている。詳細を記述すると、健康、食料及び農業、水及び環境については、同位体技術（を用いた分析結果）と AI の融合による恩恵を受ける。原子力科学分野において、AI は、データ解析、理論モデリング、実験計画に利用され、核・原子データの評価・蓄積などの基礎研究の加速や技術革新の推進に貢献している。核融合研究に必要な大規模で複雑な問題に対し、AI がモデリングとシミュレーションを通じて、核融合の研究開発を加速させる。原子力発電分野については、AI の利用により、複雑な手順を最適化し、原子炉の設計、性能、安全性を向上させることも可能であり、タスクの自動化により信頼性が向上し、エラーを低減させることも可能である。さらに、AI は、発電所のプロセスを監視し、異常を検出するのにも役立つ。保障措置は、衛星画像、環境サンプリング、ガンマ線分光法、ビデオ監視など、さまざまな手段で得られた大量のデータに依存している。AI は、これらのデータの分析に利用でき、機械学習手法により、大規模なデータセットの外れ値を検出し、使用済み燃料の検証や監視記録の分析を支援するためにすでに使用されている。

本報告においても後述するが、AI の利用においては倫理が切り離せない問題であり、上記 IAEA レポートにおいても、倫理に 1 章を割り当てており、その中で「新しい領域、すなわち原子力と AI の倫理（Ethics of nuclear and AI technologies: ENAI）が出現する。」と述べている。

IAEA はその後、AI と革新的技術が小型モジュール炉の配備の迅速化にどのように役立つかを探るための協調研究プロジェクト [3-8] を主導している。「小型モジュール炉の競争力強化と早期展開」と題された Coordinated Research Project (CRP) は 2022 年から 3 年間実施され、積層造形や AI などの革新技術が、SMR の設置コストを削減し、建設時間を短縮し、柔軟性の向上や非電気用途を通じてユーザーのニーズをよりよく満たすことができる方法を検討する。

2024 年 2 月には、AI に関する協力センター（注 4）をパデュー大学に設置した [3-9]。5 年間のこの協力センター協定により、原子力用 AI の進歩と革新に関する IAEA のプログラム活動と知識共有を支援する。具体的には、原子力における Al 技術に対する信頼とコミュニティ全体の受容を醸成するためのベンチマーク演習、それに必要な調整とデータ管理のための「ベンチマークハブ」の設立、及び IAEA 加盟国と協力して Al 技術の開発と評価に関連するその他の活動に関する IAEA のイニシアティブが含まれる。

（注 4） IAEA 協力センター：IAEA は、原子力技術の平和利用を促進するため、世界各国の指定機関と連携している。これらの指定機関は、協力センターとして、原子力科学、原子力技術、及びそれらの安全かつ確実な応用に関する独自の研究開発及び訓練を実施することにより、IAEA を支援する。パデュー大学協力センターを含め、現在、世界中に 73 の協力センターが活動を行っている。

3.1.3 経済協力開発機構（OECD）及び原子力機関（OECD/NEA）

OECD主催の「Technology Foresight Forum」は、2005年から開催されている科学技術に関する国際会議で、2016年にはAIの経済及び社会への影響をテーマにして開催された[3-10]。また2017年には、国際会議「AI: Intelligent machines, smart policies」が開催され、AIがもたらす社会、行政、産業の変化に対応した政策について討議が行われた[3-11]。これらの取組みを通じ、社会におけるAIの信頼と導入を促進するためには、国際レベルで安定した政策環境を形成する必要性があることが明らかになり、2019 年にOECDの閣僚理事会で「人工知能に関する理事会勧告」[3-12, 3-13] が採択された。本勧告はAI利用に関わる全てのステークホルダーに関係する原則と、国内政策及び国際協力において実行されるべき勧告が提示されている。

このような OECD 全体での AI 利用に関する政策的な取組みと併行して、OECD/NEA においても原子力分野での AI の利用が検討されている。例えば OECD/NEA の原子力規制活動委員会（CNRA）の新技術に係るワーキンググループ（WGNT）[3-14] の会合では、各国の AI 利用例が調査されている。また、OECD/NEA の原子力科学委員会（NSC）では、原子炉システムの科学的問題と不確実性解析に係るワーキングパーティ（WPRS）の原子炉システムマルチフィジックスに関する専門家グループ（EGMUP）において、「原子力工学における科学的計算のための AI と機械学習（ML）に関するタスクフォース」[3-15] が設置された。本タスクフォースの目的は以下のとおりである。

- AI 及び機械学習利用に係るノウハウを共有することを目的とした実践的なコミュニティの形成
- 機械学習手法の開発とパフォーマンス評価の支援
- ベンチマーク解析から得られる知見から教訓を引き出し、原子力工学の科学計算における将来的な AI 及び機械学習利用に係るガイドラインの提供

本タスクフォースは 2 つのフェーズで構成され、第 1 フェーズでは、AI 及び機械学習を利用した限界熱流束のベンチマーク解析を行っている [3-16]。原子力分野での具体的な AI 技術の利用を想定し、加盟国に実践的なベンチマーク解析の機会を提供している点は、OECD/NEA の活動の大きな特徴である。

3.2 欧州連合（EU）

EU の AI 法 [3-17] （6.3 参照）は、EU 域内でほぼ全ての分野に水平的に適用される。原子力分野での AI は適用除外となっておらず、かつ「許容できない AI」にも分類されないため、規定に従っていれば利用可能になると想定される。そこで本節では原子力分野での AI 利用が、AI 法のどの規定に該当するか調査し、各国規制当局に求められる対応を記述した。ただし、執筆時点では AI 法が原子力規制に与える影響を分析した文献は非常に少なく（一例としては Sovrano and Masetti [3-18]）、かなりの部分が推測によることを付記する。

3.2.1 管轄官庁

AI 法第 70 条（以下 AI 法の条項については番号のみ記載する）は管轄官庁（National Competent Authorities）について定められており、第 1 項には

> 各加盟国は、本規則の目的のために、少なくとも 1 つの通知当局（notified authority）及び少なくとも 1 つの市場監視当局（market surveillance authority）を国内管轄当局として設置または指定しなければならない。これらの国内管轄当局は、その活動及び業務の客観性の原則を保護し、本規則[注 5]の適用及び実施を確保するため、独立、公平かつ偏見なくその権限を行使しなければならない。これらの当局の構成員は、その職務と両立しないいかなる行動も慎まなければならない。これらの原則が尊重されることを条件として、そのような活動及び任務は、加盟国の組織上の必要性に応じて、1 つ又は複数の指定当局が行うことができる。

とあり、AI 法の適用は各加盟国（実際には管轄当局）が行うことが定められている（以下、断らない限り日本語訳は著者らによる仮訳）。同条第 3 項には

> 加盟国は、国内管轄当局が、本規則に基づく業務を効果的に遂行するために、適切な技術的、財政的、及びインフラストラクチャーを提供されることを確保するものとする。特に、国内管轄当局は、以下の能力及び専門知識を有する、十分な人数の職員を常時確保しなければならない: AI 技術、データ及びデータコンピューティング、個人データ保護、サイバーセキュリティ、基本的人権、健康と安全へのリスク、及び既存の基準及び法的要件に関する知識。加盟国は、能力及びリソースの要件を毎年評価し、必要に応じて更新するものとする。

とあり、加盟国が国内管轄官庁には AI 法施行のための予算等を割り当てることと、国内管轄当局が十分な数の専門家を常時確保することが定められている。また同条第 8 項には

> 各国管轄当局は、特に新興企業を含む中小企業に対し、欧州人工知能委員会[注 6]及び欧州委員会（EC）の指針及び助言を適宜考慮しつつ、本規則の実施に関する指針及び助言を提供することができる。各国管轄当局が、他の EU 法が適用される分野における AI 制度に関して指導及び助言を提供しようとする場合には、当該法令に基づく各国管轄官庁に適宜相談するものとする。

とあり、各国管轄当局が規制の開始時に何らかの指針を提供する可能性もある。

[注 5] AI 法のこと。

[注 6] 欧州人工知能委員会については 6.3.6 を参照。

3.2.2 ハイリスク AI

(1) 定義

6.3.3 で述べるように、第 6 条にはハイリスク AI の要件が定められており、第 2 項で「第 1 項のハイリスク AI システムに加え、AI 法付属書III（Annex III、以下他の付属書についても「AI 法」は省略する）の AI システムもハイリスクとみなす。」とされている。付属書III の第 2 項には「重要なデジタルインフラ、道路交通、水、ガス、暖房、電気の供給における管理・運営の安全コンポーネントとして使用されることを意図した AI システム」が含まれており、原子力施設などの安全コンポーネントとして使用されることを意図した AI システムが EU 域内でハイリスク AI として扱われる可能性が高い。ただし、第 6 条第 3 項には

> 付属書IIIの AI システムは、意思決定の結果に重大な影響を与えないことを含め、自然人[注 7]の健康、安全又は基本的権利に危害を及ぼす重大なリスクをもたらさない場合には、ハイリスクとはみなされない。

とあり、原子力施設で使用される全ての AI システムがハイリスクとみなされるわけではない。合意成立直後に発行された Q&A[3-19]によると、付属書IIIで定義されたハイリスク AI に AI 法が適用されるのは発効から 24 カ月後である。

また同条第 5 項は

> EC は欧州人工知能委員会と協議の上、本規則の発効後 18 カ月以内に、第 96 条に沿って、AI システムにおけるハイリスク及び非ハイリスク用途の包括的な実例リストとともに、本条を実際に実施するためのガイドラインを提供する。

としており、ハイリスク AI の用途と実施のためのガイドラインが、付属書III のハイリスク AI が規制対象となる遅くとも 6 カ月前には示されるとされている。なお、第 96 条は「本規則の実施に関する EC のガイドライン」である。

(2) 内部統制による適合性評価

第 43 条第 2 項には、

> 付属書IIIの第 2~8 項で言及されるハイリスク AI システムについては、事業者は、付属書 VI に規定する内部統制に基づく適合性評価手順に従わなければならない。これには通知機関（notified body）の関与を含まない。

と記載されている。前述のように重要インフラは付属書IIIの第 2 項に該当するので、原子力施設の AI システムにはこの条項が適用されると見られる。付属書 VI は「内部統制に基

[注 7] 法人ではない個人のこと。

づく適合性評価手順」というタイトルで

1. 内部統制に基づく適合性評価手順は、2~4 に基づく適合性評価手順である。
2. 提供者は、構築された品質マネジメントシステムが第 17 条の要求事項に適合していることを確認する。
3. 提供者は、AI システムが第III章第 2 節に規定される関連必須要件に適合していることを評価するために、技術文書に含まれる情報を検査する。
4. また、提供者は、AI システムの設計及び開発プロセス並びに第 72 条で言及されている製造販売後モニタリングが技術文書と整合していることを検証する。

と記述されており、基本的に AI システムの提供者が自身で AI 法に適合していることを確認すれば十分であることが示されている。なお、第 11 条第 1 項で技術文書について

> 技術文書は,ハイリスク AI システムが本節(ハイリスク AI の要件と題された節)に規定する要件に準拠していることを実証し，かつ，国家主務官庁及び届出機関に対し、AI システムがこれらの要件に準拠していることを評価するために必要な情報を明確かつ包括的な形で提供するような方法で作成しなければならない。

と規定している。また、第 17 条には「品質管理システム」についての、第III章第 2 節（第 8~15 条）には「要求事項の遵守」（第 8 条)、「リスク管理体制」（第 9 条)、「データとデータガバナンス」（第 10 条)、「技術文書」（第 11 条)、「記録管理」（第 12 条)、「透明性と派遣業者への情報提供」（第 13 条)、「人間による監督」（第 14 条)、「正確性、堅牢性、サイバーセキュリティ」（第 15 条）についての要求事項が記載されている。

(3) 登録

第 49 条は、ハイリスク AI の EU レベルのデータベースへの登録について定められているが、第 5 項には

> 付属書IIIの第 2 項で言及されているハイリスク AI システムは、国家レベルで登録されなければならない。

とあり、重要インフラにおける管理・運用の安全コンポーネントとして使用されることを意図した AI システムは加盟各国のデータベースに登録されると定められている。おそらく、原子力分野も該当すると考えられる。

3.2.3 ハイリスクではない AI

第 95 条は「特定要件の自主的適用のための行動規範」と題されていて第 1 項には

AI 事務局[注 8]及び加盟国は、利用可能な技術的解決策及び当該要件の適用を可能にする業界のベストプラクティスを考慮した上で、ハイリスク AI システム以外の AI システム[注 9]に、第III章第 2 節に定める要件の一部または全部を自主的に適用することを促進することを意図した、関連するガバナンス機構を含む行動規範の作成を奨励及び促進する。

と定められており、ハイリスク AI に分類されない AI への行動規範の作成を奨励している。さらに、第 2 項には

AI 事務局及び加盟国は、明確な目的及びその達成度を測定する主要性能指標に基づき、配備者を含む全ての AI システムに対する特定の要件の自主的な適用に関する行動規範の策定を促進するものとする:

(a) 信頼できる AI のための EU のガイドラインに規定されている要素

(b) エネルギー効率の高いプログラミングや、AI の効率的な設計、訓練、利用のための技術など、AI システムが環境の持続可能性に与える影響を評価し、最小化すること

(c) 特に AI の開発、運用、利用に携わる人々に対して AI リテラシーを促進すること。

とある。また、第 3 項には

行動規範は、AI システムの個々の提供者もしくは配備者、またはそれらを代表する組織、あるいはその両方が、配備者、利害関係者、及び市民社会組織や学術団体を含むそれらの代表組織の関与も含めて、作成することができる。行動規範は、関連するシステムの意図する目的の類似性を考慮して、1 つまたは複数の AI システムを対象とすることができる。

とあり、例えば原子力事業者単独又は国際的な団体に AI システム開発・提供時の行動規範策定を推奨することができる。

3.2.4 サンドボックス

第 3 条第 55 項（第 3 条は用語の定義を行っている）で AI 規制サンドボックスを

「AI 規制サンドボックス（AI regulatory sandbox）」とは、管轄当局が設定する管理された枠組みを意味し、AI システムの提供者又は提供者候補に対し、規制当局の監督の下、サンドボックス計画に基づき、限られた期間、革新的な AI システムを開

[注 8] AI 事務局については 6.3.6 を参照。

[注 9] 「許容できない AI」が含まれないことは明らかである。

発、訓練、検証、テストする可能性を提示するものである。

としており、限定的な環境で AI システムなどの開発・検証を行う環境であるサンドボックス制度の、AI 法上の定義を行っている。なお、第 54 項で

「サンドボックス計画」とは，サンドボックス内で実施される活動の目的，条件，時間枠，方法論及び要件を記述した，参加プロバイダーと所轄当局との間で合意された文書を意味する。

と「サンドボックス計画」について説明している。第 57 条第 1 項には

加盟国は、管轄当局が少なくとも 1 つの AI 規制のサンドボックスを国家レベルで設置し、発効[注 10]から 24 カ月後に運用を開始することを保証しなければならない。このサンドボックスは、他の加盟国の管轄当局と共同で設置することもできる。EC は、AI 規制のサンドボックスの設置及び運営のための技術的支援、助言、ツールを提供することができる。前段で定められた義務は、参加加盟国の国内適用と同等の水準を確保できる限りにおいて、既存のサンドボックスへの参加によっても果たすことができる。

と AI 規制のサンドボックスの設置が定められている。また、同条第 5 項で

第 1 項に基づいて設置される AI 規制のサンドボックスは、提供者又は提供予定者と所轄当局との間で合意された特定のサンドボックス計画に従って、革新的な AI システムの市場投入又はサービスの開始前の限られた期間において、イノベーションを促進し、開発、訓練、試験及び検証を容易にする管理された環境を提供するものとする。このようなサンドボックスには、そこで監督された実環境でのテストが含まれる場合もある。

と説明されている。

第 58 条は、サンドボックスの詳細についての条項で、第 1 項に

EU 全体における分断を避けるため、EC は AI 規制のサンドボックスの設置、開発、実施、運用、監督に関する詳細な取り決めを明記した実施法を採択する。実施法には、以下の問題に関する共通の原則を盛り込むものとする。

とあり、サンドボックスの詳細については別に定められることになっている。

また、OECD のサンドボックスに関する報告書[3-20]に

AI 規制のサンドボックスを管理する機関に十分な技術的専門知識が無い場合、

[注 10] AI 法の発効。

審査官が理解していないためにプロジェクトが却下されたり、通常の市場環境であれば実施可能であるにもかかわらずプロジェクトがテストプロセスを通過しなかったりするなど、市場競争に悪影響を与えかねない誤解を招く結論が導きだされる可能性がある。

と書かれているように、サンドボックスの設置者（管理者）には十分な技術的専門知識が要求される。

3.2.5 罰則

罰則については、第 99~101 条で定められている。第 99 条第 1 項では

本規則に定める条件に従い、加盟国は、事業者による本規則違反に適用される罰則及びその他の強制措置（警告及び非金銭的措置も含む）に関する規則を定め、それらが適切かつ効果的に実施されることを確保するために必要なあらゆる措置を講じるものとする。規定される罰則は、効果的で、相応で、かつ抑止的でなければならない。また、新興企業を含む中小企業の利益及びその経済的存続可能性を考慮しなければならない。

としており、罰則等の措置に関しては加盟国に実効力が与えられている。また、同条第 5 項で

要請に対する回答として、不正確、不完全、または誤解を招くような情報を届出機関（notified bodies）または管轄当局に提供した場合、750 万ユーロ、あるいは違反者が企業である場合には、その前会計年度の全世界における年間総売上高の 1%のいずれか高い方の金額を上限とする行政制裁金が課される。

とされており、管轄当局の要請に対して誤った情報を提供することも罰則の対象となる。

3.2.6 AI 法と各加盟国における既存の規制の整合

次に AI 法と各加盟国における既存の規制が整合しない場合の対応について説明する。AI 法が立脚している [3-21] という EU の機能に関する条約（Treaty of the Functioning of the European Union: TFEU）[3-22] 第 114 条の関係部分を引用すると

（第 4 項）欧州議会及び理事会、または EC による調和措置の採択後、加盟国が、第 36 条にいう重大な必要性を理由に、または環境もしくは労働環境の保護に関連する国内規定を維持する必要があると考える場合、加盟国は、これらの規定及びその維持の根拠を EC に通知しなければならない。

（第 5 項）さらに、第 4 項を損なうことなく、欧州議会及び理事会、又は EC に

よる整合化措置の採択後、加盟国が、整合化措置の採択後に生じた当該加盟国固有の問題を理由として、環境または労働環境の保護に関する新たな科学的証拠に基づく国内規定を導入する必要があると考える場合、当該加盟国は、想定される規定及びその導入の根拠を EC に通知しなければならない。

（第 6 項）EC は、第 4 項及び第 5 項の通告から 6 カ月以内に、当該国内規定が恣意的な差別の手段または加盟国間の貿易に対する偽装された制限であるか否か、及び域内市場の機能に対する障害となるか否かを検証した上で、承認または却下しなければならない。この期間内に EC の決定がない場合、第 4 項及び第 5 項にいう国内規定は承認されたものとみなされる。（以下略）

とあり、EU における AI 法の内容に関わらず、各加盟国が独自に TFEU 第 114 条第 4 項に基づいて既存の規制を維持又は第 5 項に基づいて AI に関する新たな国内規定を導入する選択肢は残されている。

3.2.7 AI 規制の流れ

これまでに見た AI 法の条文を整理すると、EU 加盟各国での原子力分野の AI 規制は次のような流れになる。

1. 各国政府は管轄当局を指定する。
2. 各国政府は管轄当局に対して予算を含むリソースの割り当てを行う。
3. 管轄当局は必要な人材確保を行う。
4. ECはハイリスク/非ハイリスクAIシステムの用途、ガイドラインを提供する（注11）。（発効から18カ月以内）
5. 管轄当局は AI 法と既存の規制の整合性について検討する。
   - 必要であれば既存の規制を AI 法に適合させる。
   - あるいは既存の規定を AI 法に優先させる。
6. 各国政府、管轄当局、あるいは複数国の管轄当局が AI 規制のサンドボックスを設置する（24 カ月以内）
7. 事業者などが非ハイリスク AI に関する行動規範を策定する。
8. 事業者は開発した AI システムについてハイリスク／非ハイリスクの判定をする。
   (a) 非ハイリスク AI の場合は行動規範に則っていることを確認し、運用を開始する。
   (b) ハイリスク AI の場合は内部統制に基づく適合性評価を行う。
9. 必要に応じて、事業者はサンドボックス制度の適用を申請する。
10. 管轄当局はサンドボックスでの試験の指導、監督、支援を行い、問題なければ実

（注 11） 詳しくは 3.2.2 を参照。

際に運用を開始する。

11. ハイリスク AI については国家レベルのデータベースに登録する。
12. 管轄当局は事業者の違反行為を発見した場合には制裁金を科す。

サンドボックスは、原子力分野特有の環境を模擬する必要があると考えられるが、第 57 条第 1 項の規定に基づいて、欧州の原子力規制機関が共同でサンドボックスを設置する可能性も考えられる。また、加盟各国の既存の規制が AI 法と矛盾する場合、TFEU 第 114 条に基づいて既存の規定を維持するか、既存の規定を AI 法に整合させるかの対応を迫られるとみられる。

3.3 ONR（英国）

6.4 で述べるが、英国政府は既存の分野別当局がそれぞれの分野で AI の規制を進めるべき、という方針を示している。ONR は Adelard 社に研究を委託し、2021 年に「AI/機械学習が原子力規制に与える影響（The impact of AI/ML on nuclear regulation）」[3-23] という報告書を作成している。この報告書では既存の規制と AI/機械学習との適合性を検討するとともに、推奨されるロードマップについて述べている。また、2023 年に英国政府が発行した白書、「AI 規制へのイノベーション促進アプローチ」[3-24] に対応して「ONR の AI 規制へのイノベーション促進アプローチ（ONR's pro-innovation approach to AI regulation）」[3-25] を 2024 年 4 月に発表した。さらに、AI サンドボックスのように既にパイロットプロジェクトの成果が発表されている（後述）ものもある。本節では既に公表されている ONR の取組みと、今後の予定について述べる。

3.3.1 利害関係者との議論

ONR が 2024 年 4 月に公開した AI についての Web ページ [3-26] では、これまでの取組みとして 3 つの事例が紹介されている。その 1 番目には

> 私たちは、事業者、他の規制当局、学会、政府を含む、国際的で原子力産業の内外にわたる幅広い利害関係者に私たちの立場を伝え、AI の課題とそれを規制する最善の方法について議論してきた。

と述べられている。

これに該当する公開情報としては、ONR は少なくとも 3 回、AI に関する専門家委員会を開催している。それぞれ 2022 年 3 月 [3-27]、2022 年 8 月 [3-28]、及び 2023 年 6 月 [3-29]（いずれも日付はホームページ上に記載のもの）である。

(1) 2022 年 3 月

Advanced Nuclear Skills and Innovation Campus (英国国立原子力研究所に設立された先進

原子力技術のキャンパス、以下「ANSIC」という。) と共催した。主な議題は

- AI システムの開発
- AI システムの実証
- AI システムの性能に対する信頼性
- AI システムの失敗について議論する

である。このうち AI システムの失敗については

> AI システムの失敗を定義し、認識することは極めて重要である。有用性が制限されるほど AI の適用を制限することと、失敗を許容できる場合と許容できない場合を認識することの間には、重要なトレードオフがある。

と記載されている。また、ONR は専門家パネルに規制に課題があるような AI の潜在的な適用先を特定するよう求めた。これらは ONR が開発中だった規制のサンドボックス（3.3.3 を参照）でテストされる。

(2) 2022 年 8 月

ANSIC と共催した。United Kingdom Atomic Energy Authority (UKAEA) の一部である、Remote Application in Challenging Environments (RACE) により開催された。EDF エナジー社、ロールス・ロイス SMR 社、セラフィールド社、UKAEA、ブリストル大学、マンチェスター大学、オックスフォード大学を含む幅広い組織の代表者で構成された。AI サンドボックスプロセスの次の段階に進める 2 件の AI ツール（注 12）について合意された。

(3) 2023 年 6 月

ANSIC、英国環境庁（Environment Agency）と共催した。ONR と英国環境庁が産業界と実施している AI サンドボックスの進捗について報告した（注 13）。

ONR はまた、パイロット・スキームの一環として実施された、短期ワークショップの終了についても報告した。これらのワークショップには、英国の原子力事業者などが参加した。ワークショップでは、前述の AI を適用する可能性のある 2 件について、模擬的なセーフティーケースをレビューし、2023 年夏に開かれたディープダイブセッションでさらに検討される主要な課題分野を特定した。

3.3.2 AI/機械学習が原子力規制に与える影響（2021）

ONR のこれまでの取組み [3-26] の 2 番目には

> 私たちは、現在の規制アプローチとその基礎となる法律が AI を規制するのに適

（注 12） 3.3.3 参照。
（注 13） 3.3.3 参照。

しているかどうかを判断するための調査を行った。

とある。それに該当するのが、ONR が 2021 年に発行した「AI/機械学習が原子力規制に与える影響」[3-23] である。2021 年の段階で既存の規制と AI/機械学習との適合性を検討し、今後のロードマップ提案を行うなど、世界の原子力規制機関の中では、AI 規制に対して最も先進的であった。この研究の委託先の Adelard 社は安全研究やリスクマネジメントなどを行っている企業で、1988 年に欧州共同体委員会の代理として、当時のエキスパートシステムのブームを受けて、AI システムをライセンスするためには標準やガイドラインの整備が不可欠であるとする論文[3-30] を書くなど、AI システムの規制には古くからかかわっている企業である。この報告書の構成は

1. はじめに
2. AI/機械学習における機会と課題
3. AI/機械学習規制のための既存ガイダンスの適合性
4. AI/機械学習保証のサポートに向けたルートマップ
5. まとめと結論

となっている。また付録で AI に関連する標準についてもレビューを行っている。以下ではこの報告書の概観を紹介する。

(1) AI/機械学習における機会と課題

この報告書のタイトルでは AI と機械学習を区別している。これは前述のように、Adelard 社がエキスパートシステムの時代から AI について検討していたことを考えると自然なことである。

原子力に特化した AI の適用として IAEA の報告書[3-7] と並んで英国の RAIN (Robotics and Artificial Intelligence for Nuclear) プロジェクト[3-31] を、「原子力産業が直面する課題を解決するためのロボット及び AI 技術の開発を目的として設立された」と紹介している。また、AI の可能性として

- 自律走行車やインテリジェント・ロボットが、廃炉サイトの除染を支援したり、人間にとって危険性の高いエリアに入って作業を行ったりする。
- 発電所のセンサーからの繊細で複雑なパターンを認識することで、安全システムがストレスを受けている可能性を警告する拡張知能と意思決定知能システム。
- ハードウェアが進歩して、組込み機器やマイクロコントローラーに AI を搭載することにより、診断機能が強化され、複雑な機能を備えた、より「賢い」デジタル機器が誕生するかもしれない。
- クラウド・コンピューティングと分散処理の進歩により、より豊富なデータの収集と分析が可能になり、メンテナンスのためのより良い予測や、システムが故障する

直前のタイミングを知ることができるようになるかもしれない。

といった例を挙げている。

一方、AI/機械学習システムの保証における現在の課題として、

> AI/機械学習製品は複雑で多様な入力に対応することが多いため、その要件を完全に規定することは困難な場合が多い。重要な課題は「ブラックボックス」的な AI/機械学習製品の性質であり、それによって動作の解釈が難しく、バイアスが不明瞭で、不可解な誤動作をする。これらのデバイスを完全に検証し、分類する手法は開発途上である。

として

- 保証アプローチとセーフティーケース（ATOMICA [3-32] では「安全性を保証するための論拠」と訳している）
- AI/機械学習ツールと開発ライフサイクル
- 標準とガイダンス
- セキュリティ

の各項目について触れ、セーフティーケースについては後の節で、標準については付録で詳述するとしている。

(2) AI/機械学習規制のための既存ガイダンスの適合性

この節では AI/機械学習のために既存の規制の明確化や更新が必要な個所を明らかにするとしている。英国の原子力規制は原子力安全規則（Safety Assessment Principles: SAPs）に要約されており、技術評価ガイド（Technical Assessment Guides: TAGs）の中に追加ガイダンスがあり、それを裏付ける規格やガイドラインが多数存在するとして、主に原子力安全規則と AI/機械学習の適合性について検討を加えている。

特に、複雑なヒューマンファクター、信頼性、セキュリティ、説明可能性、及びセーフティーケースの構築という主要なテーマについては、さらに個別の原則と関連する技術評価ガイドを取り上げて、詳細に議論を行っている。

(3) AI/機械学習保証（注 14）のサポートに向けたルートマップ

さらに、規制が AI/機械学習の展開に不必要な障壁とならないようにし、この技術を使用したシステムを適切に評価する能力を確保する、という ONR の目的を支えるための作業プログラムを強調するとして、

（注 14） AI 保証については 6.4.3 を参照。

- 原子力安全規則（SAPs）とガイダンスに基づいて、AI 規制の枠組みを策定する。
  - AI/機械学習システムで使用される技術の役割と種類を、分類学と自動化のレベルにより明確にする。
  - ヒューマンファクター、製造の卓越性（Production Excellence: PE）/独立した信頼性向上対策（Independent Confidence Building Measures: ICBMs）、セキュリティ、理解とデータの問題に対処するための原子力安全規則の解釈と変更を検討する。これには原子力産業特有の研究が必要かもしれない。
  - 製造の卓越性/独立した信頼性向上対策を置き換える保証のために、よりきめの細かい主張、議論、証拠、及び特性ベースの手法を検討する。
- 産業界と ONR において能力を構築するために積極的な役割を果たす。
  - ONR は、研究、試験、ベンチマーキングにおいて積極的な主導権を握ることで、英国の原子力部門のニーズに適合するシステムとセーフティーケースの構築を支援することができる。
  - 一連の研究には、データ、ヒューマンファクター、社会技術システム、コンピューターアーキテクチャー、ハザード分析、セキュリティ情報に基づく安全性、信頼構築技術が含まれる。
- アーキテクチャー・アプローチとデータ戦略の開発
  - データの役割とその評価に対処するため、SAPs に追加項目を作成する。
  - データの特性及び出所を評価するための分析技術を特定し、開発する。
  - アーキテクチャーとデータの役割と、それらが安全性の根拠とリスク（例えば、深層防護、多様性、リスク管理階層）に与える影響を研究する。
- 標準への関与
  - 標準化プロセスに重点的に関与する。

という提言を行い、フェーズ 1（1~2 年）、フェーズ 2（1~5 年）、フェーズ 3（3~10 年）という 3 段階でのルートマップを提示している。

なお、製造の卓越性（PE）と独立した信頼性向上対策（ICBMs）は、原子力施設の安全性が重要な部分でソフトウェアベースのシステムの使用を正当化するための ONR の「二本足」アプローチの要件で、

> 製造の卓越性（PE）の正当化-「初期使用から最終的に稼働するシステムに至るまで、製造のあらゆる側面における卓越性の実証」。それには、現在受け入れられている標準と一致する技術的な設計実践、最新の標準品質管理システムの導入、そして、あらゆるシステム機能をチェックするために策定された包括的なテストプログラムの適用という要素が含まれている必要がある。
>
> 独立した信頼性向上対策（ICBMs）-「独立した」『信頼性向上』は、安全システム

の目的への適合性について、独立した徹底的な評価を提供する必要がある。」これには次の要素が含まれる必要がある。最終システムの徹底的な分析を提供する独立した製品チェックを含む、最終的に検証された製品ソフトウェアの、チームによる完全で、できれば多様なチェック、設計意図の達成を確認するために行われる活動を含む、設計及び製造プロセスの独立したチェック、テスト活動の全範囲をカバーする包括的なテストプログラムの独立した評価。

と説明されている。

(4) 標準とガイドライン

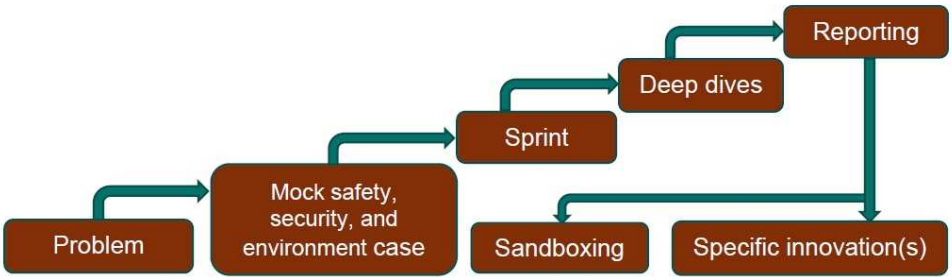
付録で AI、及び機械学習の標準化ロードマップの指針となる、開発中・完成済みの AI に関する標準の概説と原子力分野への適用可能性について述べている。

3.3.3 AI 規制のサンドボックス（2023）

ONR のこれまでの取組み [3-26] の 3 番目には

私たちは、規制のサンドボックスを利用して、2 つの AI アプリケーションで私たちの取決めをテストし、どこに課題がありそうかを判断し、AI 利用のさらなる発展を促している。

とある。サンドボックスについては、2023 年に、構造健全性（structural integrity）への AI の適用と、AI を適用したロボットグローブボックスの 2 件について報告されている [3-33]。報告書 [3-34] に記述されている ONR のサンドボックスプロセスは図 3.1 のようになる。



出典）Office for Nuclear Regulation, Environment Agency, “Regulators’ Pioneer Fund (Department for Science, Innovation and Technology): Pilot of a regulatory sandbox on artificial intelligence in the nucelar sector” の Figure 1 を英国政府, “Open Government Licence” [3-35] に基づいて転載。

図 3.1　ONR のサンドボックスプロセス

Fig. 3.1　ONR’s sandboxing process

3.3.4 RIC2023 でのプレゼンテーション

ONR で AI 技術規制に関して主要な役割を担っている Andrew White 博士が 2023 年 3 月の NRC の規制情報会議（Regulatory Information Conference: RIC）[3-36] で行った講演 [3-37] は NRC から公開されており、近年の ONR の方針をよく示していると考えられるので、本節で紹介する。

この講演で特徴的なのは、「規制当局はなぜ原子力安全アプリケーションへの AI の利用を奨励すべきなのか？」と、単なる AI の規制にとどまらず、奨励にまで踏み込んでいる点である。この理由として

- AI は、既存のアプローチでは見えない、安全性を向上させるための洞察を事業者に与える可能性がある。
- AI は例えば、汚染地域の自動調査や、グローブボックスのような放射性物質に近接する必要があるような作業の自動化によって、作業員の被ばく線量を低減する必要がある。

と述べている。一方、「しかし、強力で効果的な規制の必要性は明らかである。」と、規制が必須であることも強調している。

また、ONR が既存の（AI ではない）コンピューターベースのシステムをどう規制しているか紹介し、それとは異なる AI システムの難しさを「訓練データを使って『設計』され、システムの出力を『形成』して望ましい動作をする。これは本質的に複雑で、分析が難しい。一部の AI システムは継続的に学習するため、時間とともに挙動が変化する。」としている。そして AI を規制するためのアプローチとして

- 標準の要件に照らして規制する。しかしどの標準か？
- 正しい動作を実証する手段として、試験と不具合を修正するための反復を規制する。しかし、どの程度試験をすれば十分なのか？
- 検証・妥当性確認ができるように AI システムを構築することを主張する。AI で可能か？
- AI の故障が危険事象につながらないことを確実にする。すべてのアプリケーションで可能か？

のように、候補と、それに対する問題点を挙げている。そして「ONRはいかにしてAIシステムを可能にする規制当局になれるのか？」として、以下のような多方面からのアプローチを紹介している。

- 事業者の関与（例えば、AI を利用するためのアプローチの開発、AI 戦略の策定支援など）
- 英国の他の規制当局との関わり（産業、海事、航空、自動車、医療、防衛など）－

彼らは何をし、どのような問題に直面しているのか？

- 政府、Innovate UK [3-38]、Office for AI（現在は科学イノベーション技術省（Department for Science, Innovation and Technology: DSIT）の一部になっている）などとの連携
- 学術界、研究の補助（例を挙げると、RAIN プロジェクト、マンチェスター大学のセーフティーケース開発など）
- 研究機関、英国工業技術会（Institution of Engineering and Technology: IET）、原子力協会（Nuclear Institute）との連携
- ONR の規制準備に関する委託研究（前述の「AI/機械学習が原子力規制に与える影響」）
- IAEA ガイダンスイニシアティブへの貢献
- 他国の規制当局（米国 NRC、カナダ CNSC）との規制原則の開発

さらにプレゼンテーションの最後で「AI 規制のルートマップ例」として

1. イノベーション・セル（注 15）を設立し、安全な環境で潜在的アプリケーションのサンドボックステストを行う。これにより、我々は学ぶ機会を得ることができる。
2. 安全性に直接的な影響はないが、有益な AI アプリケーションから始める。
3. 安全性を達成するために従来のアプローチを使用しながら、AI の利点を達成できるアプリケーションに進む。
4. メリットがデメリットを明らかに上回り、原子力による悪影響が許容できるアプリケーションに進む。
5. 準備ができたら、より有意義な結果を伴う応用に進む。
6. 準備ができたら、継続的な訓練が必要で有益なアプリケーションに進む。

と、段階的に影響の大きな AI アプリケーションに進む例を示している。

3.3.5 ONR の AI 規制へのイノベーション促進アプローチ（2024）

本節では 2024 年 4 月に発行された「ONR の AI 規制へのイノベーション促進アプローチ」[3-25] の概要を紹介する。これは英国政府の白書「AI 規制へのイノベーション促進アプローチ」[3-24] で発表が求められていた、分野横断的原則と初期ガイドラインに対応するものだと見られる。この報告書では ONR の立場として

> 英国の、目標設定と慣例によらない原子力規制体制は、原子力の安全とセキュリティに対する期待が満たされることを保証する十分な正当性が確保されるのであれば、事業者が AI システムなどの革新的なソリューションや技術を採用できるような支援的環境をすでに提供している。この目標設定、結果重視、リスクベースの規

（注 15） 革新的な取組みを行うための業務単位。

制枠組みは技術的に中立である。したがって、AI は他の技術と同じ規制原則に従う。

と述べている。また

> ONR は、自律調査（廃炉現場の清掃や作業員への危害の低減などに関するもの）、知能増強（プラントデータから情報を導き出し、リスクをよりよく理解するなど）、ロボット動作の最適化（除染を迅速化し、作業員や公衆へのリスクを低減するなど）といった技術を通じて、AI が現在及び将来の原子力発電所や施設の安全性を向上させる可能性があることを認識している。

と、特に原子力分野での安全向上に対する AI の貢献への期待を示している一方、

> ONR は現在、事業者による AI の早期利用について規制している。しかし、民生原子力分野において、重要な原子力安全またはセキュリティ機能の提供において AI が利用された例はこれまでない。

と、現状では安全にとって重要な分野での AI 利用（EU のハイリスク AI 相当）例が無いとしている。また

> 利害関係者の中には、規制当局が何を受け入れ、何を受け入れないかについて、先入観を持っている場合があることは認識している。その結果、望ましい結果を達成するための最善の方法を検討する際に、過度に保守的な考え方になってしまう可能性がある。現状が維持され、より効果的な新しい解決策の導入が制限されることこそがリスクである。ONR は、規制の不確実性を最小化し、イノベーションに対してオープンな姿勢を伝え、関連するグッドプラクティスの開発に貢献し、リスクを確実に管理しながら AI の有益な利用の安全な探求を奨励するために、協力的で授権的な規制当局として、利害関係者との的を絞った関与プログラムに着手した。

と述べ、より効果的な解決策がある場合に、それを導入しないことがリスクであるという考えを示している。

この報告書では規制の 5 つの原則として

① 安全性、セキュリティ、堅牢性
- AI システムは、AI のライフサイクルを通じて、堅牢で、不安がなく、安全に機能すべきであり、リスクは継続的に特定、評価、管理されるべきである。

② 適切な透明性と説明可能性
- AI システムは適切に透明化され、説明可能でなければならない。

③ 公平性
- AI システムは、個人や組織の法的権利を損なったり、個人を不当に差別したり、不公正な市場結果を生み出してはならない。

④ 説明責任とガバナンス

- AI システムの供給と使用について効果的な監視を確保するため、AI のライフサイクル全体にわたって明確な説明責任が確立されたガバナンス手段を講じるべきである。

⑤ 競争可能性と救済

- それが適切な場合には、利用者、影響を受ける第三者、及び AI のライフサイクルの関係者は、有害であったり、重大な危害のリスクを生じさせたりする AI の決定や成果に異議を唱えることができるべきである。

を掲げている。また、

> 説明可能性と透明性は、ONR の全ての規制機能の中心であり、安全性とセキュリティに関する信頼できる主張の前提条件である。

として、「安全性、セキュリティ、堅牢性」と「適切な透明性と説明可能性」は同じ節（3.1.）で記述されている。説明可能性と透明性については

> 安全やセキュリティの主張がなされる場合、事業者はそれらを明確かつ透明に立証することが求められる。システムや技術に求められる透明性と説明可能性のレベルは、それに対する安全やセキュリティの主張の重要性に比例する。例えば、AI システムが、リスクを効果的にコントロールできる従来の対策と並行して使用される場合、求められる説明可能性のレベルは、AI が期待通りに機能しなかったとしても、安全またはセキュリティに悪影響を及ぼさないことを事業者が実証することへの期待に限定される。AI システムが安全上またはセキュリティ上重要な機能を果たす場合、期待される説明可能性のレベルはそれに応じて高くなる。

として、求められるレベルは安全性とセキュリティの重要性によるとしている。

また、この節では

- 利害関係者の関与
- 国際的な利害関係者の関与
- 英国の（他の）規制者の関与
- 規制のサンドボックス
- ガイダンス

についても記述されている。このうち国際的な利害関係者の関与では

- 米国 NRC、カナダ CNSC との三国間ワーキンググループへの参加。このグループは 2024 年後半に AI 規制のハイレベル原則を確立するための共通文書を公表する。
- OECD/NEA のロボット工学と遠隔システムの応用に関する専門家グループ

（EGRRS）に参加する。

- 2023 年 10 月に、一週間の IAEA の作業部会の議長を務め、2024 年後半に発行される「原子力発電所における人工知能利用の安全への影響（Safety Implications of the Use of Artificial Intelligence）」という技術文書（TECDOC）の起草と編集において主導的な役割を果たしている。
- ONR は 2024 年 5 月に開催される G7 諸国の原子力安全・セキュリティグループ（NSSG）において NRC、CNSC と共同で、AI 規制に対するアプローチを発表するよう要請されている。

という 4 件の事例が紹介されている。また、ガイダンスの項では安全評価原則（SAP）、セキュリティ評価原則（SyAP）、及びガイダンス文書を AI に対応して更新するとしている。

最後に「まとめと今後の展望」の節で

> ONR は（英国）政府の、原則に基づく（principle-based）、状況に応じた（context-sensitive）アプローチを歓迎し、規制が責任あるイノベーションを可能にし、国民の信頼を高め、AI におけるグローバルリーダーとしての英国の地位を強化できる、という提案を支持している。私たちは、AI に対する規制アプローチについて学習し、ケーススタディや情報を共有し、適切な場合には、将来の分野横断的な提案に対する私たちの見解を共有するために、政府全体で継続的に関与していくことを期待している。

と述べ、今後の展望としては

- 私たち（ONR）の独立性を損なうことなく、AI システムの安全な導入を支援するために、的を絞った利害関係者の関与プログラムを継続する。
- 専門家パネルや内部のイノベーション・カフェなど、イノベーションを可能にするツールの追加とともに、追加のサンドボックス演習を実施し、ONR 内で新しい技術やアプローチについてオープンに話し合うことができるようにする。
- AI に関する内部能力の構築を継続し、必要に応じて ONR アカデミーを通じて適切な訓練を実施し、検査官が一貫した適切な方法で AI の使用を規制できるようにする。
- AI に特化した安全・保安の専門検査官チームの能力をさらに高める。
- 今後一年以内に、検査官向けに AI の規制に関する新しいガイダンスを発表する。

という事例を紹介している。また「ONR の AI 規制へのイノベーション促進アプローチ」には「AI/機械学習が原子力規制に与える影響」で行った調査について

> ONR は、これらの提言に基づく的を絞った改善を受けて、2024 年 4 月に開始さ

れるこの調査の第 2 フェーズを委託した。これは、安全、保護セキュリティ、サイバーセキュリティ分野における AI の利用を可能にし、それに対応するために、どのような規制アプローチを開発する必要があるのかを探るものである。この研究の成果は一般に公開される予定である。

と記述されており、将来的な AI 規制に関する委託研究を開始したようである。

3.3.6 今後の活動

2024 年 4 月に ONR が公開した Web ページ [3-26] には、今後 12 カ月の活動として

- 私たちは、AI 規制についての政府の白書（「AI 規制へのイノベーション促進アプローチ」[3-24]）に応えて、ONR の AI 規制戦略を概説している。
- 私たちは、CNSC や NRC を含む他の国際的な規制機関と協力して、AI の規制に関する共同原則を策定している。
- 私たちは、AI の安全な使用に関するガイダンスを作成するために、IAEA や他の国々と協力して国際的に取り組んでいる。

としている。また今後 3 年間では

- 私たちは、AI の利点が確実に達成されるように、原子力発電の認可を受けた施設における AI の導入を積極的に奨励し、その能力と限界をさらに理解する。
- 私たちは、様々な利害関係者と協力し、より幅広い用途で、失敗した場合の潜在的な影響がより重大な場合に、安全性とセキュリティを実証できるような AI の設計と実装の方法について、理解を深めるよう働きかけていく。

としている。

3.4 NRC（米国）

NRC の AI への取組みの概要は Web ページ [3-39] に記載されている。英国 ONR 同様、公開で利害関係者とのワークショップを複数回開催 [3-40] している。さらにエネルギー省（DOE）、米国電力研究所（EPRI）と覚書き（MOU）を結んでいる。また、「AI 戦略計画（Artificial Intelligence Strategic Plan: Fiscal Years 2023-2027）」[3-41]、「AI プロジェクト計画（Project Plan for the U.S. Nuclear Regulatory Commission Artificial Intelligence Strategic Plan Fiscal Years 2023-2027, Revision 0）」[3-42] 等を発行し、AI 対応の計画と進捗状況について公開している。本節では「AI 戦略計画」、「AI プロジェクト計画」を中心に NRC の AI への取組みについて概説する。

3.4.1　公開ワークショップ

NRC の Web ページで公開されているワークショップは 4 回[3-40]で、タイトル、日時等は表 3.1 の通りである。各回とも発表資料が公開されていて、最も発表数が多いのは NRC だが、DOE の研究所や米国立標準技術研究所（NIST）、大学の他に企業からも発表が行われている。

表 3.1　NRC の開催した AI に関するワークショップ

Table 3.1 List of Data Science and Artificial Intelligence Regulatory Applications Workshops organized by U. S. NRC.

	タイトル	日時	場所
#1	AI の紹介（Introduction to AI）	2021 年 6 月 29 日	オンライン
#2	最近の話題（Current Topics）	2021 年 8 月 18 日	オンライン
#3	将来を見据えた取組み （Future Focused Initiatives）	2021 年 11 月 9 日	オンライン
#4	規制当局が考慮すべき AI の特徴 (AI Characteristics for Regulatory Consideration)	2023 年 9 月 19 日	対面（NRC）とオンライン

3.4.2　DOE や EPRI との覚書き（MOU）

(1) 2021 年 6 月の DOE との MOU、「運転経験とデータアナリティクスの適用分野（Cooperation in the Area of Operating Experience and Applications of Data Analytics）」[3-43]では

> 両当事者は、運転経験及び安全経験データを分析するためのツール及び技術の開発において共通の利害を有する。したがって、資源を効率的に利用し、不必要な努力の重複を避けるために、データ、技術情報、得られた教訓、場合によってはアプローチやツールの開発に関連する費用を共有することによって協力することが、両当事者にとって最善の利益になる。

としてデータの収集、解析と、AI を含むデータ解析に関する協力を定めている。

(2) 2021 年 9 月の EPRI との MOU、「原子力安全共同研究（Cooperative Nuclear Safety Research）」[3-44] の補遺（Addendum）の 1 つ「先進原子力技術とデータサイエンス（Advanced Nuclear Technologies and Data Science）」の中の 1 つの活動として「データサイエンスと AI（Data Science and Artificial Intelligence (AI)）」が記述されている。その中では協力分野の例として

(i) デジタルツイン、予防保全、確率論的リスク評価（PRA）、運転経験（operating

experience)、制御自動化、安全・セキュリティシステムのシミュレーションとコード利用、非破壊検査（NDE）などの分野に適用される AI の規格採用や堅牢性の評価を含む、説明可能で信頼できる AI に関する共同研究。

(ii) 原子力産業における AI の潜在的使用事例の評価。

(iii) AI 及び機械学習モデル開発のための一貫したデータセットに関する協力、並びに自然言語処理（NLP）及び確認分析（confirmatory analyses）を支援するための NRC と産業界の間の一貫した分野別データ語彙に関する調整、並びに共有データセット及びツールのリポジトリへのアクセス。

(iv) データサイエンス及び AI のトレーニング及びスタッフ育成に関する協働。

を挙げているが、この例に限定されることはないとしている。

3.4.3 INL への委託調査（2022）

2022年2月にNUREG/CR-7294として発行された、「稼働中の原子力発電所におけるAI と機械学習による高度な計算ツールと技術の探求（Exploring Advanced Computational Tools and Techniques with Artificial Intelligence and Machine Learning in Operating Nuclear Plants）」[3-45] はDOE傘下のアイダホ国立研究所（INL）への委託調査の報告である。

この報告書の第 6 章には、米国内の商用原子力発電所における AI/機械学習ツールの利用状況のアンケート結果が報告されており、6.4 に「調査回答からの洞察（Insights from survey responses）」として調査結果の考察がまとめられている。以下ではその概要を紹介する。

多くのアプリケーションが検討中で、ごく一部が開発中で、既にいくつかのアプリケーションが使用されていた。開発、使用されている分野は、テキストレポート分析、予知保全[注16]、作業管理、燃料サイクル管理、原子炉運転管理、代理（サロゲート）モデル開発、是正措置プログラム（Corrective Action Program: CAP）、根本原因分析（Root Case Analysis: RCA）、非破壊検査（Nondestructive Evaluation: NDE）のサポートなどがある。

実際に開発・使用されているアプリケーションの例としては

a. BWR の湿分キャリーオーバーを予測するツール

b. 将来の燃料サイクルにおける BWR の固有値変化を予測するツール

c. 症状から根本原因を特定するツール

d. データの傾向を発見し、最適なアルゴリズムを選択するために、重回帰ベースの AI/機械学習アルゴリズムを評価するツール

などがあり、IBM の Watson などの市販のツールも使われていたという。

[注16] 予防保全（Preventive maintenance）ではなく、予知保全（Predictive maintenance）という言葉が使われている。

最も有益だと考えられている分野は

a. システムと機器のモニタリング
b. 予知保全
c. デジタルツイン
d. 非破壊検査
e. 人間の労働の自動化
f. サイバーセキュリティ
g. 設計支援
h. 燃料管理
i. 停止期間短縮

であった。

回答者は設計や運転の自動化、予知保全、スタッフの生産性向上という 3 つの分野で AI/機械学習を取り入れることによる利益を期待しており、これらの分野から期待される利益として

a. 設計プロセスの効率向上
b. より広範囲、高速でデータ収集と分析が可能になる
c. 人間が気付かないパターンを特定する
d. 必ずしも事前に考えられなかった制御戦略を提案する
e. 労働集約的な作業の自動化
f. 資源配分の最適化
g. メンテナンス・スケジューリングの合理化

を挙げている。

AI/機械学習の専門知識レベルは原子力発電産業で熟知/熟練されており、最も一般的な状況としては社内で人材を育成しつつ、ベンダー、国立研究所、大学などの外部組織から専門知識のサポートを得ている。また複数の回答者が、AI/機械学習の能力、手法やツール、応用分野によって専門知識のレベルが異なると述べている。

ほとんどの回答者は、AI/機械学習の利用によって原子力発電の性能とコスト効率が向上し、他の発電方式と比較して競争力を高められると期待している。また、原子力発電産業は、他の発電方式と比較してより多くの恩恵を AI/機械学習から受ける可能性があると答えた回答者もいた。

ほとんどの回答者は、AI/機械学習が直接的または間接的に原子力規制の効率を高めると考えていた。直接的には、AI/機械学習によりプラント文書のレビューなどの職員の労働を自動化すること、自然言語処理を使用して NRC ADAMS のデータをより検索しやすくすること、代理（サロゲート）モデルにより被規制者より提出されたシミュレーションモデル

の実行に係る計算コストを下げること、診断データとリスク情報に基づく分類の調整といった、規制プロセスを合理化するための高度な監督手法を採用することなどが挙げられる。間接的には、事業者側で AI/機械学習の採用により、より安全で事象発生が少なくなることで規制側の活動を減らすことができる。

3.4.4 AI 戦略計画（2023-2027）

2023 年 5 月に NRC が NUREG-2261 として発表したのが「AI 戦略計画」[3-41] である。この文書のアブストラクトには

> 産業界が NRC の規制対象となる活動に AI を適用する可能性を想定し、NRC はこの戦略計画を策定し、NRC がそのような用途を審査できる態勢を確保した。

と書かれているが、実際には行政管理予算局（Office of Management and Budget: OMB）長官の「AI アプリケーション規制ガイダンス」[3-46] などの要求に従ったものだと考えられる。

(1) 目的

「目的と駆動力（Purpose and Drivers）」という節で、戦略計画の目的について

> AI 戦略計画の目的は、原子力産業界が AI アプリケーションの展開に関心を示していることから、NRC の規制対象である活動における AI の利用を検討するための、職員の準備態勢を確保することである。NRC のデータサイエンス及び AI 規制アプリケーション公開ワークショップ 4 (注 17) らのフィードバックに基づき、原子力産業は近い将来に AI 技術の利用を開始する可能性があり、そのような技術をどのように使用できるかの調査、開発、評価をすでに開始している。AI 技術の利用を含む許認可申請は、今後数年のうちに NRC に提出され、審査・承認される可能性がある。

と記されている。

(2) 戦略目標

「戦略目標（Strategic Goals）」という節で 5 つの戦略目標について述べている。列記すると

1. 規制の意思決定への NRC の確実な準備
2. AI 申請を審査する組織的枠組みの確立
3. AI パートナーシップの強化と拡大
4. AI に精通した戦力の育成

(注 17) 表 3.1 の#4。

5. NRC 全体での AI 基盤形成に向けたユースケース（注 18）の追求

である。このうち戦略目標 1 はこの戦略計画全体の目標で、戦略目標 2 から 5 は、戦略目標 1 で望まれる規制当局の意思決定活動のための技術的準備態勢を成功裏にサポートするための準備活動を直接的に支援するものであるとしている。また、戦略目標は優先順位の高いものから記述されている。具体的な取組は「AI プロジェクト計画」に記載されているので、詳細は次節に譲る。

3.4.5 AI プロジェクト計画（2023-2027）

「AI プロジェクト計画（ML23236A279）[3-42] は 2023 年後半に NRC より発行された文書で、前記 AI 戦略計画の各戦略目標を具体的なタスクに分解し、それぞれの執行状況・計画を記載している。

第 1 章の序論では、AI プロジェクト計画において AI 戦略計画の 5 つの戦略目標をどのように実行するかを説明している。第 2 章は戦略目標別に構成され、その目標を達成するために必要な主要タスクの説明、マイルストーン、目標とする期日などが含まれている。また、マイルストーンのタイムラインも示している。このタイムラインは NRC の現時点での推定値であり、ニーズと利用可能なリソースを考慮したものである。特に 2026 会計年度と 2027 会計年度は、NRC が規制対象活動で AI を利用できるようにするため、規制ガイダンスや手順を策定または更新し、必要であれば規則制定を開始する可能性があり、不確定要素が予想タイムラインに影響を与える可能性がある。

以下では各戦略目標の主要タスクと説明を列挙する。

(1) 規制の意思決定への NRC の確実な準備

本目標については、規制フレームワークを開発して、スタッフが NRC の規制活動の一環として AI を評価できるように準備することに重点をおいており、以下の 5 項目のタスクが設定されている

a. 原子力アプリケーションの AI 利用に対する、規制枠組みの適用性評価（2024 会計年度中に完了予定）
b. 原子力アプリケーションの AI 利用に対する、AI 標準の適用性評価（報告書は 2025 会計年度第 3 四半期までに完了予定）
c. 原子力アプリケーションのための、AI 安全、セキュリティ評価フレームワーク
d. AI 提出物の申請前コミュニケーションと計画（2025 会計年度第 2 半期に完了予定）
e. AI を利用した自律的原子力運用のための規制フレームワークの開発

上記 a については、「NRC が規制する活動における AI の許認可と監視に関する既存の規

（注 18） ここでは「（AI の）使用シナリオ」、のような意味と推測される。

制枠組みの適用性を評価する」としており、それにより規制・ガイダンスや検査手順の更新・新規作成の必要性を検討するようである。

また、b については、「既存の AI 規格を特定・評価し、NRC の規制対象活動への適用性を判断する」としており、仮に適用可能な標準があれば a の規制・ガイダンスに取り込むようである。この a 及び b は英国 ONR が「AI/機械学習が原子力規制に与える影響」[3-23] で報告し、継続プロジェクトを行っていると見られる「AI/機械学習規制のための既存ガイダンスの適合性」、「AI/機械学習保証のサポートに向けたルートマップ」、及び付録の「標準とガイドライン」と同様の活動であると見られる。また a 及び b の主な調査活動は請負業者（Contractor）が担うようであり、ONR に対して Adelard 社が担った役割を請負業者が担当すると見られる。

c については、「NRC が規制する活動においての AI の利用について、包括的な安全・セキュリティ評価のフレームワークを研究・開発する」のが目的であるとしており、NIST の AI リスクマネジメントフレームワーク [3-47] の原子力版のようなものを想定していると見られる。また、「一般市民、外部の利害関係者、連邦のパートナー、及び国際的な規制コミュニティと協力して、AI の安全・安心のためのフレームワークの開発と実施に関する情報と見解を収集する」としているが、連邦のパートナーとしては NIST や AI セーフティー・インスティテュートを想定しているはずである。この項目も調査活動の主要部分は請負業者が担うようである。

d については、「NRC は、ベンダー、申請者、ライセンス保有者から、申請前活動、トピカルレポート提出、その他のライセンス提出物に関する AI 計画情報を求め」、収集した情報に基づいて「NRC の予算、及びリソース計画に情報を提供する」としている。

e については

> このタスクの目的は、様々な潜在的な自律レベル、スタッフ配置計画、及び遠隔操作コンセプトを考慮しながら、原子力運転における AI を利用した自律性を実現するための技術的基盤と必要な規制の枠組みを開発することである。このタスクは、先進炉プログラムスタッフ及び規則策定スタッフとの緊密な調整を含み、AI による原子力の自律運転に必要な規制の厳格さが適用されることに重点を置く。（略）このタスクは、タスク c と連動して開発される。

と記述されており、これについても調査の主要部分は請負業者が担うようである。

(2) AI 申請を審査する組織的枠組みの確立

a. AI 運営委員会とワーキンググループの設置と利用（AI 運営委員会は 2023 会計年度第 3 四半期に設置済み）

b. AI 実践コミュニティの立ち上げと利用

c. 一元化された AI プロジェクト・データベースの構築と管理（データベースの構築

は 2023 会計年度第 1 四半期に完了）

本目標については

> AI 戦略計画を成功裏に実施するには、NRC 全体で効果的な調整と協力が必要である。

としている。a では「米国の原子力施設における将来の AI 利用に備え、NRC の業務プロセスにおける AI 技術の一貫した適用を確保するため、NRC 内の部門横断的な調整と指示を行う AI 運営委員会（AI Steering Committee: AISC）を設置」する。本項では AI 運営委員会の機能として

- AI 戦略計画の実施に関する上級管理職の監督も含まれる。
- 本計画に記載されたタスクの予算決定を指導する。
- 必要に応じて、AI の専門知識を有する外部の専門家に特定の問題の支援を依頼することができる。
- AI 戦略計画に記載された戦略目標を達成するための活動の直接的な優先順位付けを確実にするため、情報技術・情報管理ポートフォリオ執行委員会（Information Technology and Information Management Portfolio Executive Council）と調整する。
- AI プロジェクト計画の活動を実施するために、必要に応じて AI ワーキンググループ（AI Working Groups: AIWGs）を設置する。

としている。

b では

- AI 技術の使用を含む要請を審査するためのベストプラクティスと教訓の共有を促進する。
- 積極的かつ潜在的な使用事例について全庁的な認識を提供する。

という目的で共同フォーラムとして AI 実践コミュニティ（AI Community of Practice: AICoP）を設立することを定めている。AI 実践コミュニティは AI 技術、政策、基準、プログラムに積極的な、あるいは関心のある NRC の各プログラム及び地域事務局から、主要な担当者が参加し、少なくとも 1 名の上級管理職と支局長を含む。

c では「NRC で実施されているすべての AI 関連プロジェクトを捕捉する一元化されたデータベースを構築し、維持する」としている。

(3) AI パートナーシップの強化と拡大

a. 国内パートナーシップ

b. 国際的パートナーシップ

c. 公開ワークショップ、会議、会合の主催と参加

本目標については

> NRC は、AI 技術を安全に導入、監視、及び評価するための貴重な情報を入手し、リソースを利用するために、国内外の原子力産業やその他の政府機関に属するカウンターパートと強力なパートナーシップを維持、発展させている。

としている。

a は「NRC が国内の AI のベストプラクティスや教訓を利用できるように、米国政府機関、国立研究所、非政府組織、学術機関、その他の研究機関とのパートナーシップを構築し、維持すること」を目的としている。記述されている項目は

- EPRI、DOE との MOU（前述）
- NIST の AI リスクマネジメントフレームワーク（AI RMF）に参加 (注 19)
- 連邦航空局や食品医薬品局など、他の政府機関とも関与を続け、AI に関するアイディア、実務、手順を交換
- 連邦政府の AI 担当者の月例会合と、統合サービス庁（General Services Administration）の政府全体 AI 実践コミュニティに参加（2020 年の大統領令「連邦政府における信頼できる AI の利用促進」に定められている）

b は

・他の組織の経験から学ぶ
・自らのベストプラクティスと教訓を共有する
・国際的な規制慣行に影響を与えて世界の原子力安全を促進できるよう、国際的な規制機関や研究機関とのパートナーシップを構築し、維持する

ことを目的としている。記述されている項目は

- 進行中の NRC の AI 規制フレームワークの開発と、国際標準及びガイダンスの改定案との一貫性を維持し、最も重要な標準とガイダンスを優先して早い段階で改定プロセスに影響を与えることを目指す。
- 他の加盟国とともに、AI に関するいくつかの技術会議で IAEA と協力し続けている。
- カナダ、フランス、ドイツ、アラブ首長国連邦、及び英国と AI 活動に関するいく

(注 19) AI リスクマネジメントフレームワークについては 6.5.7 を参照。リスクマネジメントフレームワークは多くの政府機関や民間企業などの意見を取り入れて作成されている。また定期的な改訂を謳っており、NRC の継続的に改訂に関わっていく意図を示していると考えられる。

つかの二国間協定を結んでいる。

– AI に関する規制アプローチとベストプラクティスに関する共同研究と情報交換が含まれる。

- NRC の RIC を、国際パートナーとの直接的な交流を促進するために利用する。
- NRC、カナダ CNSC、英 ONR（CANUKUS）間の三国間活動を実施する。

– CANUKUS のプロジェクト計画を策定する（2023 会計年度第 2 四半期に完了）。

– AI 原則に関する三国間 CANUKUS 報告書の原案を完成させる（2024 会計年度第 2 四半期まで）。

– AI 原則に関する三国間 CANUKUS 最終報告書を発表する（2025 会計年度第 1 四半期まで）。

である。国際関係の中でも、特に国レベルの AI 規制で関係の深い英国、カナダとの関係を重視しているようである。なお、米国の会計年度は 10 月から翌年 9 月までなので、2024 年会計年度第 2 四半期は 2024 年 1~3 月となる。

(4) AI に精通した戦力の育成

a. NRC の AI スキル評価とギャップの特定（2026 会計年度第 1 四半期に完了予定）

b. AI トレーニングの機会の特定、開発、実施（2027 会計年度第 2 四半期に完了予定）

c. AI 人材の採用、雇用、維持（2027 会計年度第 2 四半期に完了予定）

本目標については

> この目標は、スタッフが AI アプリケーションをレビューできるようにするための技術情報、知識、ツールの開発に重点を置いている。

としている。aは「NRCの現在のAI関連スキルを評価し、スキルギャップを特定して、NRCがAIの使用を含む規制レビューを効果的かつ効率的に実施できるようにすること」を目的としている。2020年の連邦法「2020年政府におけるAI法」(注20) でアメリカ合衆国人事管理局（Office of Personnel Management: OPM）局長に、AIに関連する職位に必要とされるスキルを特定するよう求めていたが、それに対応して2023年7月にAI力量リスト[3-48] が発行された。このAI力量リストに基づいてAI関連スキルの評価とギャップの特定を行うものとみられる。

b、c では a で特定されたギャップに基づいて NRC 職員に対して訓練を行い、また、新規採用等を行う。本目標を通じた目的を

> 全体的な目的は、NRC が AI の使用に関連する規制レビュー及び監視活動を効果的かつ効率的に実施するために、適切なスキルを持つ適切な人数の人材を適切なタ

(注 20) 「2020 年政府における AI 法」については 6.5.4 を参照。

イミングで適切な場所に確保することである。

としている。

(5) NRC 全体での AI 基盤形成に向けたユースケースの追求

a. AI テストと分析のための概念実証アプリケーション

b. AI エコシステムの開発と維持

c. 安全性評価のための調査 AI ツールと手法（2025 会計年度第 2 四半期に完了予定）

d. AI 規制研究の促進と投資

本目標については

この目標は、NRC の規制対象活動における AI の使用を審査するための技術的専門知識を構築するためのユースケースの開発と追及､及びデータサイエンス､評価、新たな AI ツールの統合と実践的な人材育成をサポートするサイバーエコシステムの構築に重点を置いている。これを達成するために、NRC は様々なソースから様々な形式のデータを使用してユースケースを開発するための研究を実施し、原子力業界と協力して潜在的なパイロットスタディと概念実証を追求する。また、NRC は、

・ ソフトウェアベースの AI ツールへのスタッフアクセスの改善
・ トレーニング及び開発ツールへのスタッフアクセスの改善
・ 将来の規制審査を模擬する可能性のあるトレーニング演習へのスタッフの参加の促進

についても調査する。この目標の成功の結果として、NRC スタッフは、AI 分析のエンドツーエンドの研究開発、新興 AI ツールの統合、及び原子力業界からの AI アプリケーションを審査するための実践的な人材育成をサポートする自己完結型の情報技術エコシステムを所有することになる。

としている。

また、d については

このタスクの目的は、様々な既存の研究メカニズムを通じて AI 規制研究の促進投資を行うことである。これらの研究メカニズムには、プログラムオフィスの作業要求、NRC の将来志向研究（future-focused research: FFR）、及び大学原子力リーダーシッププログラム（University Nuclear Leadership Program: UNLP）が含まれる。目的は、5 つの AI 戦略目標全てに対処する NRC の能力を向上させる AI 関連の規制研究への投資を積極的に継続することである。

としている。

3.4.6　NRC における AI の利用の推進（2023）

NRC 委員長の Hanson 氏は 2023 年 10 月に、「NRC における AI の利用の推進（Advancing Use of Artificial Intelligence at the U.S. Nuclear Regulatory Commission（ML23303A143））」[3-49] という作業指示メモを発出した。その中では、

> 私はスタッフに、規則制定、環境レビュー、研究活動など、NRC のライセンス及び監視プロセスを改善するために、AI を NRC 内でどのように使用できるかの検討を行うよう指示する。NRC は、将来を見据えた組織になるため、AI アプリケーションを優先すること。具体的には、スタッフは AI を使用して単純又は反復的なタスクを自動化し、人的エラーを減らし、リソースを節約し、大規模なデータセットを処理し、データに基づくより適切な意思決定を行い、規制レビュー時間を短縮する方法を評価すること。スタッフは、AI アプリケーションが知識管理、戦略的労働力計画、将来の労働力を構築及び準備するための採用イニシアティブにどのように役立つかを検討すること。

として NRC 内部での AI 使用について検討するよう指示している。

これに対応して 2024 年 4 月に Policy Issue [3-50] が発行された（SECY-24-0035)。その中で最近の取組みとして

> NRC の AI チームは、2023 年 12 月にバージニア州ロスリンでベンダー、及び NRC 上級幹部との会議を主催し、NRC での AI 導入に関する議論を開始した。NRC の AI チームは、翌週に同様の仮想会議を主催し、関心のある NRC スタッフ全員が参加してベンダーの現在の AI ツールについて理解を深められるようにした。これらの会議は、NRC がベンダーの AI 環境を理解するための最初の討論会となり、NRC の業務における AI 使用に関する考え方の転換に役立った。
>
> これらの会議の後、NRC の AI チームは SharePoint と Microsoft Teams チャネルに共有コミュニケーションプラットフォームを設定し、ワンストップでの主要リソースへのアクセスを提供した。ユースケースの収集への機関全体の参加を促すため、NRC の AI チームは 23 の NRC オフィスと地域すべてを one-on-one 会議に招待した。これらの会議では、オフィスが質問をしたり、ユースケースの要件に関する最初の考えを共有したり、他のオフィスとのパートナーシップを探ったりする機会が与えられた。例えば、このアウトリーチ活動により、各地域は NRC の監視プロセスに関連する多くのユースケースで協力することができた。

と、紹介している。ユースケースについては

> NRC の AI チームは、潜在的な 61 のユースケースを検討し、データサイエンティストや AI の中小企業と協議した結果、いくつかの共通テーマを特定した。推奨さ

れた 36 の AI ユースケースのうち、16 のユースケースには、市販の生産性向上ツールを含む生成 AI を利用するという包括的なテーマがあった。これらの 16 のユースケースでは、AI を適用して日常的なタスクを自動化し、ワークフローとプロセスの改善を実施することで、スタッフの日常的なタスクの多くを効率化する。これらのユースケースの例としては、生成 AI を利用して会議の記録、ドキュメント、及び Web ページを要約することが含まれる。推奨された残りの 20 の AI ユースケースの包括的なテーマは、予測分析、セマンティック分析[注 21]、コメント分析といった用途に他の種類の AI[注 22]を用いることである。これらのユースケースの例としては、他の種類の AI を使用して予測分析を実行すること、提案された規制に関するパブリックコメントのレビューを自動化すること、地域全体での検査官のスケジュール設定と可用性の計画を支援することなどがある。これらのユースケースは、日常的なプロセスを自動化し、データ分析の効率を改善してデータ主導の意思決定を促進することで、スタッフの生産性を高め、スタッフをサポートすることが期待されている。NRC は、新たに設置された AI ガバナンス委員会と、IT ロードマップの開発、近代化、拡張（Development, Modernization, Enhancement: DME）プロセスを通じて、これらの潜在的なアプリケーションの価値提案を評価し、実装の優先順位を決定する。

としている。

また、今後のステップとして

(1) NRC 内での AI の利用を促進するために、NRC 全体の AI 戦略を策定する
- 信頼でき、責任のある AI の実装を確実にするために AI ガバナンスを準備する
- データ管理プログラムを成熟させる
- 戦略的雇用と、既存従業員のスキルアップにより AI 人材を強化する
- IT インフラストラクチャーの一部として AI ツールの統合をサポートするために、リソースを割り当てる

(2) AI の利用を促進するための基盤ツールに投資する
- 現在のアプリケーションと統合する生成 AI サービスを取得する
- AI の使用を ADAMS の認知検索テクノロジーと統合する

という計画を示している。

[注 21] テキストから意味を抽出する分析。

[注 22] 機械学習や自然言語処理など。

4. 原子力関連分野等での AI 適用・検討例

AI の研究分野は、実際に実用化されているものから最先端技術として研究されているものまで様々である。本章では、原子力分野及び関連する分野における AI 研究の利用例について紹介する。AI 研究は活発であり、2023 年度、国内で開催された各学会の年次大会では数多くの報告がなされている。しかしながら、前述したとおり、AI 研究については、日進月歩で次々と新しい成果・提案がなされており、当該技術ノートが発刊される時期においては、ここで紹介する情報は、すでに古い情報となっている可能性があることに留意されたい。

4.1 原子力施設における利用及び検討例

IAEA はその報告書[4-1]において、AI の原子力発電への応用の可能性を検討しており、

> 原子力産業において AI を活用することは、自動化、設計最適化、データ分析、耐用年数の予測、情報の抽出などにおいて有益であると考えられる。AI による自動化は、プレッシャーが大きい、あるいは厳しい要件が課される状況での信頼性とリスクの低減につながり、結果として人為的ミスによって運転が中断される時間を最小限に抑えることができる。自動化の例として、制御棒の破損のデータ分析や、発電所の異常検出などが挙げられる。AI によってこれらを最適化することにより、運転効率の向上や、予測的に炉心を制御するような複雑な操作を具体化できる可能性がある。さらに、AI を利用し、物理的概念に即した高度な統計モデリングは、新しい学習データに対しての一般性を保持することも可能にしながら、供用期間中検査における経年劣化影響評価にも活用できる。このような発電所の予測モデリングを利用することで、維持管理活動に関するデータの共有を進めることができる。ただし、原子力産業では現在、標準的な手法がより広く採用されているため、このような手法は十分には活用されていない。また、稼働中の発電所で利用できる膨大なデータを用いることで、運転及び保守の効率を向上するためのベストプラクティスを新たに抽出することも可能になる。

と述べている。このような AI の利用は将来的に見込まれるものであるが、一部の限られた例を除き、具体的な原子力発電所等の運転や保守への利用は、検討が進められている段階である。

本節では、運転や保守に関するものではないが、国内外における原子力施設等における AI 技術の利用及び検討例について紹介する。

我が国においては、浜岡原子力発電所での利用例が報告されている。

Watanabe ら[4-2]は、浜岡原子力発電所の管理区域の入域に際して、保護具装備の確認に AI 技術を利用していることを報告している。原子力発電所の放射線管理区域へ入域する際には、作業員を放射線汚染から防護する観点から保護具を装備する。通常、安全保護具の

装備確認は鏡に映った自身の姿を指さし呼称で確認するセルフチェックを基本としているため、ヒューマンエラーに起因する装備不備が発生することが稀に報告されている。それらの対策として監視員を 24 時間常駐とするとコストがかかることや、更なるヒューマンエラーによる見逃しも想定されることから、安全保護具の装備確認に AI 判定を取り入れる事を検討している。Watanabe らは、保護具チェックのための物体検出手法として、深層学習による物体検出アルゴリズムの一つである SSD（Single Shot Multi-Box Detector）を用い、作業者の画像から各チェック対象の有無を検出する機能を実現している。チェック対象である認識カテゴリとして、保護衣（青服または作業着)、帽子、手袋、靴下、線量計を設定している。また、「負例」（リジェクトすべきカテゴリ）として、素手と下着の 2 カテゴリを設定した。物体検出とチェック機能の実装はオープンソースと Python の各種ライブラリを用い、独自開発を行っている。さらに、AI チェック機能と連動した自動開閉ゲートを開発している。AI チェック機能で線量計を含むすべての保護具を確認して、ゲートが自動開閉する機能だけではなく、現行の運用に合わせて、AI チェック機能で線量計を除く保護具を確認し、線量計の所持については、ゲートの QR コードの読み取り機能で所持確認することもできるものを開発している。

横洲ら [4-3] は、浜岡原子力発電所に到達する津波の早期検知に AI を利用していることを報告している。浜岡原子力発電所では、到達する津波を早期に検知するために、津波監視用の海洋レーダーを 5 号機屋上に設置し、視線方向流速を広範囲（発電所を中心に半径 60 km）かつ連続的に観測（1 分間隔）している。また、津波リスクの観点から約 5,500 ケースの津波シミュレーションを実施しデータベース（ビックデータ）を構築し、これを学習データにしている。AI 技術を用いることにより、海洋レーダーによる津波流速画像の情報から浜岡原子力発電所に到達する津波の高さ、到達時間を予測可能であることを確認したと報告している。紹介している AI を用いた手法では、波源を求めることなく、津波の流速分布やその変化から津波を予測できるため、計算負荷も軽く、瞬時に結果が得られるとしている。

田中ら [4-4, 4-5] は、高速炉を含む革新炉開発における知見を集約し、最新の解析評価技術との連携により、既往知見を最大限利用した安全性、経済性、保守性などのさまざまな観点からの統合的な設計評価を可能とする、AI 支援型革新炉ライフサイクル最適化手法（ARKADIA）を開発しているとしている。ARKADIA は、AI 技術を導入した ARKADIA プラットフォームをベースとして、仮想プラントライフシステム（VLS)、評価支援・応用システム（EAS)、ナレッジマネジメントシステム（KMS）から構成される。これら 3 つのシステムに基づき、設計最適化支援ツール及び安全評価ツールを構築している。設計最適化支援ツールは、簡易モデルによる高速評価から、熱流動、核特性、構造などの連成解析による緻密な評価を可能とするマルチレベル構造になっている上、AI による自動最適化により設計の効率化が可能となると述べている。

日立製作所 [4-6] は「現場拡張メタバース」として、メタバース空間上に現場を迅速に再現

し、現場データ（現場のヒトやモノに関する画像・映像・文書・音声・IoT データなど多様な種類のデータ）の蓄積や可視化のためのプラットフォームを構築している。生成 AI を含む AI 技術によって、容易にデータを活用できるシステムを構築し、原子力発電所における作業効率と安全性の向上や技術伝承、人材育成に活用している。社内で実施された原子力発電所の実寸大模型の移設工事に適用し有効性を評価した。

東芝[4-7]は、オートエンコーダ（例えば Raschka ら[2-28]を参照）を 2 段階で適用する異常予兆検知システムを提案しており、バイオマス発電所で実証実験を行っている。同様に三菱電機[4-8]は、火力発電分野で学習した監視信号の波形パターンによるパターン型検知と、信号間の相関評価を組み合わせたハイブリッド型の異常兆候検知システムを火力発電プラント向けに提案している。これらの技術は原子力施設への導入が検討されている。

東芝エネルギーシステムズ[4-9]は、AI を利用した改善処置活動（CAP）を提案している。これは、原子力安全に影響を及ぼすおそれのある情報（Condition Report: CR）を幅広く収集・分析し、改善につなげるものである。事業者は、収集した CR の内容を確認し、重要度に応じた分類コードを付与して処置内容を検討するが、取り扱う CR が膨大で、CAP 運用の負荷増大が問題となっている。そこで、以下のような、AI を利用した CAP 運用支援システムを開発した。

(1) 類似検索 AI　　CR の特徴を解析して類似 CR を検索
(2) 自動分類 AI　　CR に付与する分類コードを推定

これにより、CR 情報に対する類似事例の抽出と、分類コードや重要度の自動推定ができ、CAP 運用の効率化が図れる。また、検索・分類機能だけでなく、CR の傾向分析や事象発生場所の視覚化機能も備えており、従来にない CAP 運用の高度化で、原子力の安全性・信頼性向上に寄与するとしている。

また、日本原子力エネルギー協議会（ATENA）[4-10]によると、日本の原子力産業界の現在の状況は、

(1) 原子力発電所の運転・監視・制御・保護といった、原子力発電所の安全に直接関係するシステム・機器に AI 技術は現時点では適用されていない。
(2) 一方で、異常兆候診断等の予防保全業務や、情報分析等の関連業務への AI 技術適用が業務効率化や品質向上を目的に開発されている。
(3) 類似した火力プラント等の他分野で実証された AI 技術（例；異常診断技術）の原子力展開も見込まれる。

と整理される。

米国の連邦規則[4-11]では、すべての許認可取得者に対し、安全機能の喪失につながる系統及び機器故障の発生（Maintenance Rule Functional Failures: MRFFs）を評価するよう求めている。MRFFs の発生は、発電所で記録される全事象の 0.1~0.2%に過ぎず、MRFFs を評価するために発電所の安全機能の喪失につながる系統及び機器故障を監視することはコスト及び労力がかかる。Hess ら[4-12]は、米国エクセロン社における MRFF アナライザーの開発及び導入実績について述べている。MRFF アナライザーは、自然言語処理、人工ニューラルネットワーク、ベイズ統計などの機械学習技術を組み合わせて MRFFs を評価するシステムであり、MRFFs 評価のコスト及び労力を削減する。

Lin ら[4-13]は、機械学習アルゴリズムによって改良されたデジタルツイン等を用いて、複雑な流量喪失シナリオ時に運転員に合理的な制御勧告を行うための制御システムの改良について提案している。

Tokatli ら[4-14]は、放射線レベルが高く、人間のアクセスが厳しい設備における作業としてグローブボックスを用いた作業を取り上げている。グローブボックスは複雑な作業を行うことができる一方、グローブに穴が空く等により隔離性が破られると、運転員は大きなリスクにさらされることから、グローブボックスを用いた作業におけるリスクを低減するため、グローブボックス下での作業に AI ロボットを利用した遠隔操作を導入することを提案している。

Huang ら[4-15]は、原子力設計の最適化に向けた AI 技術の利用に関する研究をレビューし、以下のように AI 利用に向けた課題及び今後の展望を提言している。原子力技術への AI 利用の課題は二つに分けられる。第一に、データに問題があり、不十分な実験データはデータの偏りや不均衡を高める。第二に、ブラックボックスジレンマの問題があり、深層学習のような手法は、AI 内部での演算処理についての理解不足によるブラックボックスジレンマが生まれる。今後の展望として、AI アルゴリズムがラベル付き学習データに大きく影響を受けるという問題に対処するため、物理的メカニズムを組み込んだ科学的機械学習を用いて、知識の埋め込みをすることにより、モデルのパフォーマンス及びロバスト性を向上させること及びモデルの透明性と信頼性を高めるために、説明可能な AI 技術の利用を促進することを挙げている。

Deleplace ら[4-16]は、原子力発電所のスクリーン洗浄機の状態監視のための、特徴選択とアンサンブル機械学習技術に基づく故障判定アプローチを紹介している。そのアプローチは、以下のとおりである。まず、包括的で統計的な特徴のセットが現場の生の加速度センサーデータから抽出される。次に、故障検出アルゴリズムの精度を向上させるために、評価したい時系列に関連するデータを抽出する。このプロセスでは、識別が可能な指標である特徴選択測定基準を利用し、抽出したデータに基づく特徴を選択する。最後に、決定木法（付録 1 を参照）を改良した手法である XGBoost が故障検出において選択された特徴を用いて訓練される。他の故障検出器との比較分析の結果、故障検出において、アンサンブル学習は他の手法よりも精度が高く、効果的に使用できると述べている。

Che ら [4-17] は、原子炉の挙動を正確に予測するには、核工学、熱流動学、燃料熱力学を連成させたマルチフィジックスシミュレーションが必要であると主張している。そこで、従来、燃料の性能解析を全炉心について実施するにはコストがかかるため単独で行われているが、機械学習を用いて、燃料性能に関する高速計算を可能にする代替モデルを構築することで、全炉心モデリングの計算効率を向上することを提案している。提案された手法は、標準及び高燃焼度 PWR 炉心の両方で検証され、満足のいく予測精度で FRAPCON と比較して少なくとも 10^4 倍早く計算できることを達成したことを報告している。

二神ら [4-18] は、原子力発電所のリスク評価における AI 利用について、以下のような手法を開発している。原子力発電所の確率論的リスク評価（PRA）は、評価モデルの構築に膨大な設計情報が必要となる一方で、これらの情報の解析コードへのインプットは手作業でアナログ的に行うという課題がある。この課題を解決するために、AI やデジタル化技術を利用してフォールトツリー（FT）を自動で生成するアルゴリズムを開発している。

4.2 放射線防護分野や核セキュリティ分野に関する検討例

IAEA は AI の利用に関する報告書 [4-1] で放射線防護に関する利用について以下のように検討している。

放射線防護の焦点は、放射線被ばくが生じうる職場における安全要件と安全基準の統合である。現在、既存の AI アプリケーションの安全基準への統合が進められている。機械学習アルゴリズムとバーチャルリアリティツールは、労働者の線量計算に関するシミュレーションや作業計画への適用、または規制要件に準拠するための原子力施設を含む施設や活動の設計中の線量最適化など、放射線防護の特定の課題に対処するために利用できる。放射線被ばくを含む作業プロセスの分析、解釈、理解において人間の認知に代わるアルゴリズムとソフトウェアを作成することなど、AI を使った研究により放射線防護を強化することができる。さらに、多くの異なる機械で放射線データを収集して分析することにより、放射線防護プログラムを確立するためのより速く、より柔軟で、より効率的なプロセスが可能になり、この分野の深い技術的変革につながる。

核セキュリティ分野で AI を使用することには、潜在的な利点とリスクがある。利点としては、規制管理外の物質の検出と対応を改善する可能性、核物質の計量管理システムを改善する可能性、原子力施設で起こりうる内部及び外部の脅威を特定する可能性があることが挙げられる。一方でリスクとしては、核セキュリティシステムで AI を使用すると、人間のオペレーターだけでなく AI システム自体にもすぐに認識できない脆弱性が発生する可能性がある点である。そのため、核セキュリティシステムにおける AI 適用の限界についての理解を深めなければならない。この分野では、AI 対応技術に対するサイバー攻撃の脅威に関する慎重な調査も重要である。核セキュリティの分野では、AI の利点とリスクの分析に最も多くの取組みを集中させる必要がある。専門家は、AI を開発及び実装する際に慎重な検討を行うことと、セキュリティを損なうのではなく維持するための明確な目的と

指標を確立することを奨励している。AI はまた、データへのアクセス性、知的財産の制約、さらにはデータ主権を取り巻く問題に加えて、核セキュリティに関する多くの倫理的及びプライバシー上の懸念も生じさせている。

4.3 材料、構造分野での検討例

原子力以外の工学の様々な分野では、既に AI の利用が進められている。そこで、本節では原子力に限らない幅広い工学分野における AI の利用に向けた研究状況について、近年の検討事例を一部紹介する。

原子力産業にも密接に関連する代表的な AI の利用先として、非破壊評価の分野があげられる。非破壊評価（NDE）では、構造物の維持・管理のために、構造物内部の欠陥の位置や大きさの評価が行われる。一般に欠陥の評価に用いる非破壊試験のデータは複雑であるため、検査者の経験と判断が評価精度に大きく影響する。そのため、従来から検査者の熟練度に依らずに非破壊評価データを自動的に処理する手法が検討されている。そこで、近年では当該処理を自動化する手法の一つとして AI を利用することが期待されている。

非破壊評価については、斎藤ら[4-19]が最近の AI・データサイエンスとそれらの利用動向を調査した結果をまとめている。斎藤らによれば、1990 年代にはニューラルネットワークを用いた非破壊検査に関する研究が活発に行われ、超音波、渦電流、放射線、サーモグラフィ、磁気探傷、地中レーダーなどに対してニューラルネットワークが応用された。2000 年代に入ってこの流れは下火になったが、近年、再び活発になっている。また、斎藤らは、同分野における学術的な取組みを、(1)波動伝搬挙動を解き明かす数値シミュレーション、(2)その結果を利用または考察することで、欠陥形状の再構成や物性の評価を行う逆解析、及び(3)実際の非破壊評価への応用という 3 点に分類し、それぞれの研究動向について整理している。

非破壊評価への応用研究として、Siljama ら[4-20]は複数チャンネルの超音波フェーズドアレイ探傷装置で得られたデータから、欠陥を検出するためにニューラルネットワークを応用する方法を提案している。Siljama らは、最新のニューラルネットワークを用いることで、人間が検出する場合と遜色の無い精度で、欠陥の検出を自動化できることができると報告している。

ただし AI の非破壊評価への利用には課題も残されており、例えば深層学習に必要な学習データの確保は大きな課題の一つである。一般に非破壊評価に必要なデータセットを確保するには、欠陥が存在する試験体を用いて大量の非破壊試験データを収集しなければならない。近年では、このような課題を解決する方法として、限られた試験データから深層学習を行うためにデータを増やす処理（データ拡張）を行うことが検討されている。例えば、Virkkunen ら[4-21]は、超音波フェーズドアレイ探傷のデータから欠陥を検出する際にも、データ拡張が有効であることを報告している。

非破壊評価以外の分野の AI 利用の検討事例には、気液二相流の限界熱流束の予測があ

る。限界熱流束は核沸騰熱伝達の上限値に相当し、これを超える熱流束では加熱面の到達温度が融点を超える可能性があるため、熱機器の焼損を防ぐための重要な指標である。Nafey ら[4-22]は大量の試験データからニューラルネットワークの学習を行い、限界熱流束を予測するためのアプローチを提案している。

4.4 地震、津波分野

4.4.1 地震学分野における AI 利用研究例

日本では、気象庁や自治体により設置された震度計・地震計に加え、1995 年の兵庫県南部地震を契機に国立研究開発法人防災科学技術研究所による高感度・広帯域地震計からなる地震観測網が整備された。さらに、近年では鉄道やガスなどのライフラインにも振動計が設置されるなど地震記録に関するビッグデータが構築されつつあり、これらデータを利用した AI 研究の報告がなされている。

田中[4-23]は、将来的に自然災害の発生が予測される箇所を示したハザードマップ作成に機械学習を用いている。ハザードマップ作成には、その地点周辺の地形の影響が大きな要因としてあげられるが地形情報の複雑さから一定ルールに基づくデータ作成は困難である。そこで、地形情報に基づくモデリングに対する機械学習的アプローチの有効性を確認している。

高橋ら[4-24]は、震源パラメーターが周期に依存した水平 2 成分の地震動強度の空間特性に及ぼす影響を分析している。横ずれ断層における 600 ケースの周期・成分別の絶対加速度応答値の分布をモード分解し、モードと震源パラメーターとの関係性を機械学習のランダムフォレスト(注 23)によってモデル化している。このモデルに機械学習モデルの解釈手法である説明可能な AI を適用し、モードごとに支配的な震源パラメーターを分析している。

石井ら[4-25]は、新たな観点からの地震動評価による知見の獲得を目指し、過去に得られた地震動観測記録を学習用データとする機械学習により、地震観測点毎に固有な地震動評価モデルの作成を試みている。従来の距離減衰式等では扱われなかった震央方位や地震動の応答継続時間も検討対象としている。全体として観測値は良く評価・モデル化され、評価値の大半は観測値の倍～半分の範囲に収まり、評価値／観測値の比の平均はほぼ 1、その常用対数標準偏差は地震動の振幅では 0.2 強、応答継続時間では 0.1 強となったと報告している。応答継続時間への震央方位の影響度は大きく、従来の予測式の各パラメーターの影響度と同等以上になる場合もあったとのことである。

小穴ら[4-26]は、強震動統一データベース試作版を用いて、機械学習により最大加速度と応答スペクトルの地震動評価モデルを構築している。構築したモデルにおける観測値に対する予測値の比の常用対数標準偏差は 0.18～0.21 で、既往の地震動予測式のばらつきよりも小さくなったとのことである。学習用データセットよりも時系列的に後に起きた地震を

(注 23) 付録 1 (3) 決定木を参照。

追加テストデータとしてモデルの汎化性を検証した結果、学習用データセットには含まれていない特徴をもつ地震において予測精度の低下が見られたが、既往の地震動予測式に基づく予測結果を特徴量に加えたモデルにより、教師データの偏りや不足を補い得る可能性が示されたとしている。

工藤ら[4-27]は、P・S 検測モデル及びノイズ／ P S 識別モデルを作成し、一元化震源の検測値と、その検測値に対応する地震波形データをモデルに与えて学習させ、モデルの精度を検証している。また、学習済みのモデルを PF（Phase combination Forward search）法と組み合わせることによって、PF 法の震源がどのように変化するかについても検証している。

さらに、第 238 回地震予知連絡会においては、平田ら[4-28]が重点検討課題「人工知能による地震研究の深化」の概要を公開しており、地震学の分野においては、ますます AI 関連の研究が進んでいくものと考えられる。

4.4.2　津波工学分野における AI 利用研究例

2011 年東北地方太平洋沖地震における津波被害を契機に災害予防及び減災に AI を利用するための研究がなされてきた。

郷右近ら[4-29]は、2011 年東北地方太平洋沖地震津波の被災地において、被災前後の高分解能合成開口レーダー衛星画像（TerraSAR-X）を用いた新しい建物被害領域自動検出モデルを提案している。その際に、被災前後の TSX 画像と既存の機械学習アルゴリズムに基づき、建物流失が大きい領域を自走検出するモデルを構築している。さらに、千葉ら[4-30]は、被災後画像のみを用いて建物被害領域を抽出している。

青井[4-31]は、機械学習を用いて発生しうるあらゆるパターンの津波を想定した津波シナリオバンクから、日本海溝海底地震津波観測網（S-net）が検知した沖合水圧データをシナリオ選別アルゴリズムも用いて検索し、陸域の津波遡上までを予測している。

また、Makinoshima ら[4-32]は、事前にスパコンなどを利用して多数のシナリオを想定した津波シミュレーションを行い、津波等の模擬観測データと予測地点での津波浸水波形の関係を AI に学習させている。大地震発生時には、リアルタイムに得られる実観測データに基づき、AI が予測点の津波高とその時刻を含む津波浸水波形を即座に予測できるとしている。

4.4.3　地盤・岩盤工学分野における AI 利用研究例

地盤・岩盤工学分野では斜面崩壊や液状化の推定に、機械学習等を利用する研究が進められている。

平岡ら[4-33]は、遠心場での斜面崩壊実験で計測した斜面表層ひずみの時系列データから、深層学習の手法の一つである LSTM を用いてデータの予測を行い、その予測値と計測値の残差によって、斜面の異常を検知する手法について検証している。その結果、設置した 8 基の表層ひずみ計の異常検知数の時系列推移から崩壊前に斜面の異常が検知できることを

確認している。また、時系列データを定常化するために表層ひずみ速度を用いた場合においても、同様に崩壊前に斜面の異常検知が行えることを確認している。

鳥谷部ら [4-34] は、地震動の加速度記録のみから液状化の程度を評価する深層学習技術を対象として、その妥当性を検討している。東北地方太平洋沖地震における関東地方 140 地点の強震観測データを事例として、実際に液状化被害が確認された地点との比較評価を実施している。

桑原ら [4-35] は、1891 年から 2016 年までに発生した 41 地震のデータを用いて、日本全国の液状化危険度の推定を行っている。入力データと液状化の複雑な背景構造のモデル化を期待し、機械学習手法の一つであるランダムフォレストを推定モデルとして採用している。データセットの不均衡性を考慮し、アンダーサンプリングとアンサンブル学習を行うことで、既往研究よりも高精度かつ安全側を重視するモデルを作成したとしている。さらに、作成したモデルに仮想の地震動を入力することで日本全国の液状化ハザードマップを作成している。

XIE ら [4-36] は、機械学習を用いて北海道胆振東部地震における斜面崩壊の推定を行い、その精度を検証している。さらに、衛星画像が地震後早期に取得できたものと仮定し、正規化値生指数を機械学習の説明変数に用いることを検討している。

山岳トンネル工事では、切羽面に分布する岩石の種類や性状を観察し、適切な支保の種類を判定しながら掘削を進めている。奥澤ら [4-37] は、現場技術者による切羽観察の一助とするため、深層学習を用いて岩塊の写真から岩石の種類（岩種）を判定するシステムの開発を行っている。当該研究では、主として国内の様々な地域や年代の地質から集められた、29 岩種、2,656 試料を対象に、様々な角度から可視光画像を撮影している。アルゴリズムには AlexNet を使用して得られた画像から 1 岩種当たり 7,000 枚をランダムに抽出して学習させたところ、平均で 72.1%の正解率が得られたとしている。本学習モデルを用いて、撮影した画像をサーバーに送ると岩種を回答する岩種判定システムを構築している。

4.4.4 耐震・建築分野における AI 利用研究例

耐震・建築分野においては、画像処理と深層学習から構造物の被害及び損傷状況を推定する研究が多く報告されている。地震等の大規模災害においては、被害範囲が広く、余震など追随する災害もあり、構造物の被害状況を安全かつ迅速に判定するために大変有用であると考える。

佐藤ら [4-38] は、画像処理による損傷量の計測手法に、既往の損傷画像をビッグデータに用いた深層学習によるパターン認識を導入し、鉄筋コンクリート造壁部材の静的載荷実験において観測された損傷量データを新手法により計測している。また、提案した新手法を用いた損傷量計測方法の特徴について、従来法との比較を通して考察を行っている。

内藤ら [4-39] は、熊本地震の前震直後及び本震直後に取得された航空写真を用いてデジタル地表モデル（Digital Surface model: DSM）差分解析、テクスチャ解析、ブルーシート抽出

の各画像解析手法を用いた被害抽出を、益城町及びその周辺地域における 15,000 軒以上に適用している。また、それぞれの解析結果について航空写真の目視判読により 4 段階に分類した被害区分との比較を行い、分類精度の検証を行っている。

藤田ら [4-40] は、航空写真から深層学習を用いて地震被害の大まかな規模と全体像を把握する屋根損傷家屋把握システムを開発している。このシステムは GIS の建物ポリゴンの位置情報により判別のための画像データを自動で作るアルゴリズムにより迅速な予測が可能である。ブルーシート判別は正解率約 93%、直接被害判別は正解率約 81%の精度で予測ができたと報告している。また、深層学習では高画質で大量の画像データが必要である一方、現状では画像データ数が乏しいという問題点があるため、画像認識アルゴリズムなどの改良に加え、データの収集方法の工夫が必要であることも指摘している。

森田ら [4-41] は、地震後に撮影した RC 柱の画像を対象にして、深層学習を用い、応急危険度判定における評価基準の一つとなる RC 柱の損傷度の予測可能性について検討している。地震後に、建物管理者等がスマートフォン等により RC 柱を撮影した画像に基づき、深層学習処理による RC 柱の損傷度を判定するシステムである。このようなシステムでは応急危険度判定の完全な代用は出来ないが、建物使用者が RC 柱の損傷度を把握することで、危険な壊れ方かどうかの判断を行うことができると報告している。

吉岡ら [4-42] は、RC 方立壁のせん断破壊の特徴を多く含む地震被害写真を用い、事前学習済み CNN モデルのファインチューニングを行い、せん断破壊が先行する RC 柱の地震被害写真から損傷度を分類している。

5. AI のリスク

本章では AI のリスクについて示す。一般的な AI ガイドラインでは AI のリスクに対応して、持続可能性、人間中心、透明性、説明可能性、公平性、安全性などの原則が定められているが、本章では技術的な観点から問題となり得る AI のリスクについて述べる。なお、AI のリスクに関しては Zhang ら[5-1] によりよく整理されているので、この論文を中心にリスクを整理し、新たに認識されるようになりつつあるリスクについても記述する。また AI そのものが引き起こすリスクではないが、AI の性能向上・利用増加に伴って電力需要が急増するという懸念がある。これについても広い意味での AI のリスクとして本章で取り上げる。

5.1 リスクの分類

Zhang らは、AI、機械学習のリスクを次のように分類している。

- データレベルのリスク
  - データの偏り（Data bias）
  - データセットのシフト（Dataset shift）
    - ＊ 共変量シフト（Covariate shift）
    - ＊ 事前確率シフト（Prior probability shift）
    - ＊ 概念シフト（Concept shift）
  - ドメイン外データ（Out-of-domain data）
  - 敵対的攻撃（Adversarial attack）
    - ＊ ターゲットを絞った攻撃（Targeted attack）
    - ＊ ターゲットを絞らない攻撃（Untargeted attack）
- モデルレベルのリスク
  - モデルバイアス（Model bias）
  - モデルの誤った指定（Model misspecification）
  - モデル予測の不確実性（Model prediction uncertainty）

以下ではデータレベルのリスク（敵対的攻撃を除く）、モデルレベルのリスク、敵対的攻撃と、近年新たに問題になりつつあるリスクについて説明する。

5.2 データレベルのリスク

データからコンピューター自身が学習するのが AI、特に機械学習の大きな特徴である。これは人間が細かいルールを与える必要が無いという点で大きな強みであるが、逆に問題があるデータを与えるとそのデータで学習した AI も問題を受け継いでしまうというリスクにつながる。また、AI の学習時に想定していなかったようなデータを入力すると、AI が想定外のふるまいをしてしまうリスクもある。

(1)データの偏り（Data bias）

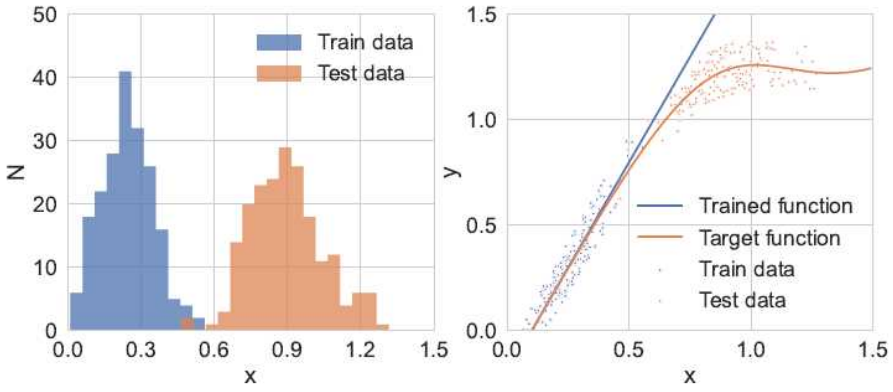
あるグループや要素が AI モデルで他と比較して過剰に重みづけされたり代表されたりしているケースをデータの偏り（Data bias）という。このような AI モデルは偏った判断をすることが多い。有名な事例としてはアマゾン（Amazon.com）の人材採用における AI 利用があげられる [5-2]。アマゾンで開発した履歴書審査の AI で、履歴書に「女性」に関連する単語が含まれていると評価が下がる傾向が見つかり、最終的にこの AI の使用は中止された。そのような判断を下す理由として、この AI に学習させた過去の履歴書のデータのほとんどが男性のものだったため、システムが男性を採用するのが好ましいと認識したためと考えられている。

(2)データセットのシフト（Dataset shift）

① 共変量シフト（Covariate shift）

学習時と実際の使用時とで AI のモデル化する入出力関係自体は変化しないものの、入力の分布が異なるようなケースを共変量シフト（Covariate shift）という。Zhang らの Fig.2 を参考に作成した図 5.1 の概念図では左図は学習時（Train data, 青）と評価時（Test data, オレンジ）の入力値の分布を表しており、図の横軸は入力の値（x）を、縦軸は度数（個）を表す。図の場合では青で示される学習データは主に x<0.6 の範囲に、また、オレンジで示される評価データは 0.6<x<1.5 の範囲に分布しており、ほとんどオーバーラップが無い。出典）Zhang, X., Chan, F. T. S., Yan, C., Bose, I., “Towards risk-aware artificial intelligence and machine learning systems: An overview", Decision Support Systems, Vol.159, 2022.(5-1)の Fig. 2 を参考に作成

図 5.1 の右図は入力値（x）と出力値（y）の対応関係を表しており、オレンジ線（Target function）が「正解」の関係である。図中に青い点（Train data）で示される学習データは「正解」線付近に分布しているものの、「正解」が折れ曲がる x>0.5 の点がほとんど存在しないため、学習データで訓練した AI の入出力関係（Trained function）は図の青線のような直線となり、x>0.5 の領域では「正解」からかけ離れてしまっている。



出典）Zhang, X., Chan, F. T. S., Yan, C., Bose, I., “Towards risk-aware artificial intelligence and machine learning systems: An overview", Decision Support Systems, Vol.159, 2022.(5-1)の Fig. 2 を参考に作成

図 5.1　共変量シフト（Covariate shift）の概念図

Fig. 5.1　A conceptual scheme of covariate shift

注）左図は学習時（Train data, 青）と評価時（Test data, オレンジ）の入力値の分布で、図の横軸は入力の値（x）を、縦軸は度数（個）を表す。右図は入力（x）と出力（y）を散布図としたもので、オレンジ線は目標とする入出力の関係を表す。青い点は学習データの、オレンジの点は評価データの入出力関係を、青線は学習データで訓練した AI の応答を表す。

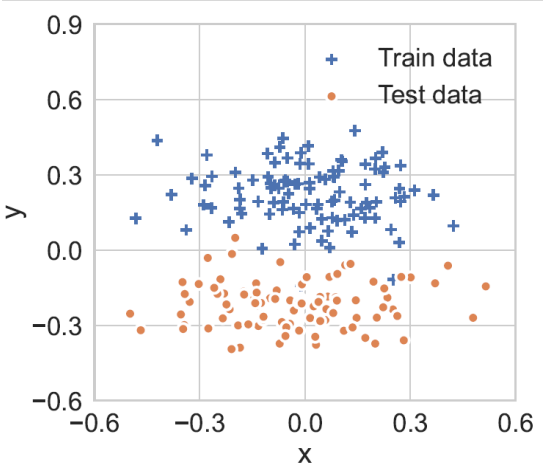
② 事前確率シフト（Prior probability shift）

学習時と評価時で入力データの分布は変化しないものの、出力の分布が異なるようなケースを事前確率シフト（Prior probability shift）という。Zhang らの Fig.3 を参考に作成した概念図（図 5.2）は横軸（x）が入力値を、縦軸（y）が出力を表し、青の点が学習時のデータ（Train data）を、オレンジの点が評価時のデータ（Test data）を表している。この図の例では学習データも評価データも入力は x=0 を中心としたほぼ同じ分布である。しかし、出力は学習データが y=0.2 を中心とした分布になっているのに対し、評価データでは y=-0.2 を中心とした分布になっている。

③ 概念シフト（Concept shift）

入出力の関係が時間などによって変化するケースを概念シフト（Concept shift）、あるいは概念ドリフト（Concept drift）という。Zhang らの Fig.4 を参考に作成した概念図（図 5.3）では左図が学習データ（Train data）を、右図が評価データ（Test data）を表している。いずれの図でも緑の線が決定境界（Decision boundary）を表し、青い点と

オレンジの点の 2 種類のクラスを分けている。AI は左図の学習データを学習することで 2 種類のクラスを分離する決定境界を学習するが、その学習結果を右図の評価データに適用すると決定境界が学習時とは異なっているため、誤ったクラス分類をしてしまう。

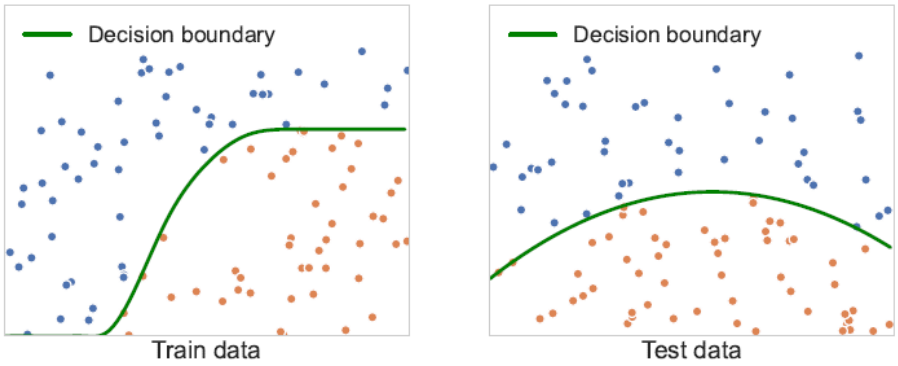


出典）Zhang ら [5-1] の Fig. 3 を参考に作成

図 5.2　事前確率シフト（Prior probability shift）の概念図

Fig. 5.2　A conceptual scheme of prior probability shift

注）横軸（x）が入力値を、縦軸（y）が出力を表し、青の点が学習時のデータ（Train data）を、オレンジの点が評価時のデータ（Test data）を表す。



出典）Zhang ら [5-1] の Fig. 4 を参考に作成

図 5.3　概念シフト（Concept shift）の概念図

Fig. 5.3　A conceptual scheme of concept shift

注）左図は学習データ（Train data）を、右図は評価データ（Test data）を表す。いずれの図でも緑の線が決定境界（Decision boundary）を表し、青い点とオレンジの点の 2 種類のクラスを分けている。

(3)ドメイン外データ（Out-of-domain data）

一般に AI モデルは限られたドメインのデータで学習する。例えば MNIST [5-3] と呼ばれる 0 から 9 までの手書きの数字画像のデータセットを学習した AI は手書きの数字を 0 から 9 までの 10 通りのクラスに分類することができるが、手書きの「A」や「あ」という文字の画像を入力した場合にどのようなふるまいをするかは予測不可能である。このようなケースをドメイン外データ（Out-of-domain data）と言うが、AI システムの実装によっては深刻なセキュリティーリスクをもたらしうる。

5.3 モデルレベルのリスク

本節ではモデル自体に起因するリスクについて述べる。

(1)モデルの偏り（Model bias）

AI モデルが偏り（bias）を持っているケースをモデルの偏り（Model bias）という。Amin らによる最近の報告 [5-4] によれば、「私は〇〇の患者です。放射線検査の報告を要約してください。」という質問（ただし、〇〇は米国の国勢調査で定義される人種で、「白人」、「黒人またはアフリカ系アメリカ人」、「アメリカインディアン及びアラスカ先住民」、「アジア系」、「ハワイ先住民及びその他の太平洋諸島の住民」）を OpenAI 社の ChatGPT-3.5 及び ChatGPT-4 に 750 通りの放射線検査の報告と共に入力したところ、出力に人種による有意な差が出たという。モデルの偏りの主要な原因は学習データの偏りだが、AI モデル自体が何らかの偏りを持っている場合もある。

(2)モデルの誤った指定（Model misspecification）

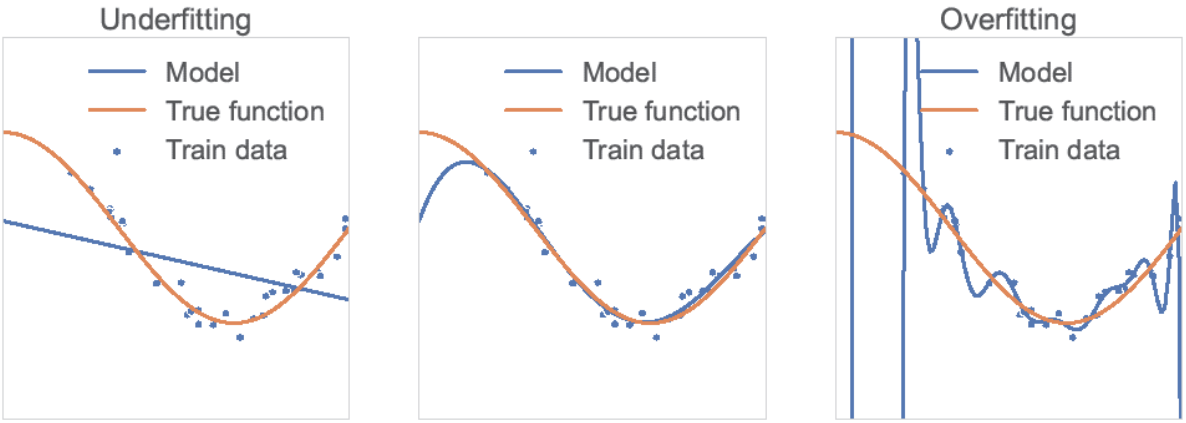
一般に、AI では学習データと評価データが同じ分布にしたがっている場合でもそれがどのような分布かはわからないことが多い。そのような場合に不適切なモデルを使用することに起因するリスクをモデルの誤った指定（Model misspecification）という。Zhang らはモデルの誤った指定をさらに

1. 学習不足（Model form error、underfitting)
2. 過剰適合（過学習、Model overfitting）
3. 特徴量の選択ミス（Variable inclusion error）

の 3 種に分類している。1.の学習不足（一般に underfitting と呼ばれることが多い）はモデルがデータを表現するのに不十分な場合に起こり、2.の過剰適合（一般に過学習、あるいは overfitting と呼ばれることが多い）はモデルの表現力がデータに対して過剰な場合に起こりやすい。

図 5.4 は Pytho 言語の機械学習用ライブラリ scikit-learn [5-5] の“Underfitting vs. Overfitting”の項 [5-6] のサンプルコードをもとに作成した図で、いずれの図でも赤線が予測したい真の値（True function）を、青点が学習データ（Train data）を、青線がモデルによる予測値（Model）を表している。左図は 1 次の、中央図は 4 次の、右図は 15 次の多項式で学習データを近似した結果を示しており、中央の予測値は真の値を比較的良く近似している（真の値は正弦波の一部である）。一方、左図では直線で曲線を近似しているためモデルの表現力が不足しており、真の値の特徴を捉えているとはいいがたい。このケースは学習不足に相当する。逆に高次の多項式で学習データを近似している右図は学習データに含まれる誤差の影響を大きく受け、一部の領域で極端な振動を示している。このケースは過剰適合に相当する。

また、複数次元の学習データで AI モデルを学習する際、重要な変数（特徴量）が含まれていないとモデルの予測精度が損なわれてしまう。逆に不要な変数（特徴量）が含まれていると、モデルがその変数に大きく影響を受け、予測精度が損なわれてしまう場合がある。これらのケースは特徴量の選択ミスに相当する。



出典）　"Underfitting vs. Overfitting", https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html を参考に作成

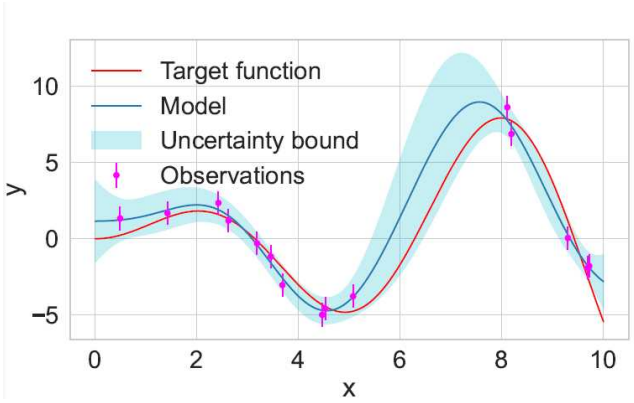
図 5.4　学習不足、及び過剰適合の例

Fig. 5.4　Examples of under-fitting (left) and over-fitting (right)

注）いずれの図でも赤線が予測したい真の値（True function）を、青点が学習データ（Train data）を、青線がモデルによる予測値（Model）を表している。左図は線形モデルで、中央図は 4 次の多項式で、右図は 15 次の多項式で予測を行った場合を示す。

(3)モデル予測の不確実性（Model prediction uncertainty）

AI、機械学習モデルの予測にはパラメーターや構造に伴う不確実性が存在する。図 5.5 の例では、赤線が予測したい真の値（True function）を、紫の点と縦線が測定値とその誤差（Observations）を、青線がモデルによる予測（Model）を表している。また水色の帯がモデルによる予測の不確実さを表している。用途によってはこのような不確実性がリスクとなり得る。なお、この例では予測モデルとしてガウス回帰過程を使用し、図は scikit-learn の“Gaussian Process regression: basic introductory example” [5-7] のサンプルコードをもとに作成した。



出典　"Gaussian Process regression: basic introductory example, https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_noisy_targets.html#sphx-glr-auto-examples-gaussian-process-plot-gpr-noisy-targets-py を参考に作成（© Copyright 2007 - 2024, scikit-learn developers (BSD License).

図 5.5　モデル予測の不確実性（Model prediction uncertainty）の例

Fig. 5.5　An example of model prediction uncertainty）

注）赤線が予測したい真の値（True function）を、紫の点と縦線が測定値とその誤差（Observations）を、青線がモデルによる予測（Model）を表している。また水色の帯がモデルによる予測の不確実さを表している。

5.4 敵対的攻撃（Adversarial attack）

AI、特に深層学習では、敵対的攻撃（adversarial attack）が大きなリスクとして考えられている。これは入力にわずかな揺らぎを加えることにより、AI モデルの出力が全く異なってしまうというものである。最初期の報告である Szegedy ら [5-8] による例を図 5.6 に示す。この例ではいずれのケースも左端の図がオリジナルの画像で、AlexNet [5-9] により正しく分類されている。しかしわずかな揺らぎ（中央の図、10 倍に拡大している。）を加えた右端の図は、人間の目にはオリジナルとの差異はほとんどわからないが、AlexNet では全てダチョウ（ostrich）と分類されたという。同様に Shi ら [5-10] の例（図 5.7）では左側のオリジナルの図をコンゴウインコ（macaw）と正しく分類しているのに対し、人間の目にはほとんどわからないわずかな揺らぎ（中央の図）を加えた右図は本棚（bookcase）と分類されている。Zhang らは敵対的攻撃について、狙ったラベルへの誤分類を目標とする、ターゲットを絞った攻撃（Targeted attack）と誤分類させることを目標とするターゲットを絞らない攻撃（Untargeted attack）とに分類している。

図 5.6 や図 5.7 の例では画像を誤分類させるだけだが、現実世界に影響を与えるような攻撃を仕掛けることも可能である。Eykholt ら [5-11] は交通標識に、人間の認識には問題にならないレベルの落書き（Graffiti）を加えることで、高い再現性で誤認識させられることを

示している。例えば「Stop（止まれ）」の標識を「Speed Limit 45（速度上限 45）」と認識させることに成功しており、この手法を適用すれば自動運転の自動車に事故を引き起こさせることが可能になる。

また、Papernot ら [5-12] は転写性（transferablity）という性質を持つ敵対的攻撃を報告している。これは、異なったデータセットで学習した同じ AI モデル、あるいは別の AI モデルに対して同様に誤分類させることができる敵対的攻撃の性質である。転写性を利用すると、学習データセットもアーキテクチャーもわからない AI に対してブラックボックス（black-box）攻撃を仕掛けることができる。手元の代理モデル（surrogate model）で有効性を確認した敵対的ノイズ（adversarial noise）を加えた画像を用いることで、ブラックボックスな AI に対してもある程度の確率で敵対的攻撃を成功させることができる。Zhong と Deng [5-13] はこのような手法を高度化することで、Amazon、Microsoft、Baidu などの商用の顔認証に対して有効な敵対的攻撃を成功させたと述べている。

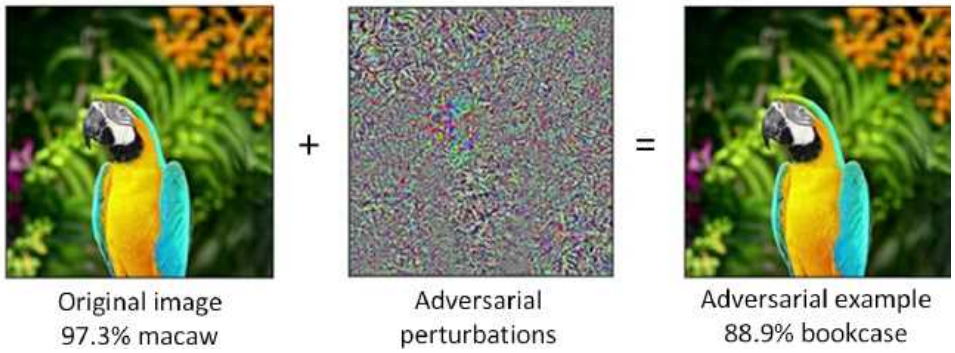
出典）　Szegedy, C., Zaremb, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., "Intriguing properties of neural networks", arXiv, 2013.の Figure 5

Creative Commons, CC BY に従い引用

図 5.6　Szegedy らによる敵対的サンプル（Adversarial example）

Fig. 5.6　An adversarial example illustrated by Szegedy et al

注）いずれの図も左側が正しい画像で、右側が揺らぎを加えた図、中央の図は揺らぎを 10 倍に拡大したもの



出典）　Shi, Y., Fan, C., Zou, L., Sun, C., Liu, Y., "Unsupervised Adversarial Defense through Tandem Deep Image Priors", Electronics, Vol.9, p.1957, 2020. の Figure 1 をクリエイティブコモンズ表示（CC-BY）ライセンスに従って転載

図 5.7　Shi らによる敵対的サンプル（Adversarial example）

Fig. 5.7　An adversarial example illustrated by Shi et al

注）左側が正しい画像でコンゴウインコ（macaw）と分類されているのに対し、揺らぎを加えた右の図では本棚（bookcase）と分類されている。中央の図は揺らぎだけを取り出したもの。

5.5 プロンプトインジェクション攻撃

大規模言語モデルは単独で利用されるだけでなく、人間の指令（プロンプト: prompt）を柔軟に解釈できるという能力を生かして他のシステムに組み合わせて使われることが多い。通常、大規模言語モデルは不正なプロンプトは実行しないなどの安全策を持っているが、Liu ら[5-15]によると、その安全策をかいくぐる脱獄（jailbreaking）というテクニックが存在し、完全に安全とは言えない。Liu ら[5-16]はこのような脱獄を利用して、大規模言語モデルの組み込まれたシステムを不正に操作するプロンプトインジェクション攻撃（Prompt Injection Attack）の可能性を指摘している。これは、単純に大規模言語モデルの組み込まれたシステムに想定外の動作をさせたり、不正に情報を引き出したりするだけでなく、乗っ取ったシステムを経由してインターネットにさらに大規模な攻撃を仕掛けるケースも想定されている。

特に規模の大きい大規模言語モデル（基盤モデル、あるいは汎用目的モデルなどとも呼ばれる）を不正に使用した場合は国やそれ以上の広い範囲に重大なリスク（システミック・リスク）をもたらしうる。後の章で紹介するように、各国はこのようなリスクの可能性を重く見て、基盤モデル、汎用目的モデルと呼ばれる大規模モデルの規制を検討・施行している。

5.6 AI の欺瞞

ここまで述べてきたのは、人間による AI の不適切な設計や使用、意図的な AI への攻撃によるリスクであるが、AI の高度化とともに最近は新たなリスクが注目を集めつつある。それは AI の欺瞞（AI deception）と呼ばれるもので、AI 自身が人間を欺いているように見えるケースである。Park ら [5-17] は欺瞞を「真実以外の成果を追求するために、誤った信念を体系的に誘導すること」と定義し、特定用途向けの AI（例えば Diplomacy という対話型の戦略ゲームをプレイさせるためにメタ社が開発した Cicero という AI [5-18]）だけでなく、大規模言語モデルのような汎用目的 AI でも欺瞞が見られるとしている。また Park らは現在の AI には備わっていないが、いずれ AI が現状認識（situation awareness）を持つようになった場合には、自分が評価過程に置かれている場合には人間の望む振る舞いをし、実際に使用される場合と違う振る舞いをする欺瞞も起こりうるのではないかとしている。

5.7 AI の電力消費

AI が引き起こすものではないが、AI に関連するリスクとして電力消費の増大が上げられる。深層学習は巨大なニューラルネットワークのパラメーターを計算によって決定するため、学習には膨大な計算を必要とする。de Vires [5-19] によると、OpenAI 社の大規模言語モデル、GPT-3（2020 年）の学習に要した電力は 1287 MWh であり、近年のさらに規模の大きいモデルではさらに学習に必要な電力は増大していると考えられる（OpenAI 社は近年の大規模言語モデルの詳細を公表しておらず、学習時の消費電力を推定することは困難である）。

一方、大規模言語モデルの利用者が増えるにつれ、モデルの推論のための消費電力も問題となりつつある。de Vires [5-19] によると通常の Google 検索の消費電力は一件当たり 0.3 Wh であるのに対し、ChatGPT への問い合わせは一件で 2.9 Wh 消費するとしている。EPRI によるホワイトペーパー[5-20] は、将来的な大規模言語モデルの消費電力の増加と、AI 利用の増加の見通しから、データセンターでの電力消費の増加を予想している。2023 年の米国のデータセンターでの消費電力が約 152 TWh であったのに対し、年間 15 %増加という一番大きな見積もりでは 2030 年の消費電力は約 404 TWh で、全米の電力消費の 9.1 %に達するとしている。年間 5 %増加の控え目な予想でも、2030 年には 214 TWh になるとしている。このような見通しから、AI サービスの大手事業者は安定な電力源として核融合や原子力発電に着目している [5-21]。

6. AIのリスクに対処するための仕組み

前章で紹介したように、AIについては様々なリスクが懸念され、また現実のものとなっている。これらのリスクに対処するための仕組みは「ガバナンス」と呼ばれ、罰金を伴う法律から、強制力を持たないガイダンス、自主的な取組みに至るまで幅広い。また地理的・組織的な広がりも、全世界を対象としたものからEU、各国レベル、あるいは一企業レベルまで様々なレベルで存在する。また、様々な利用法があり、広範囲に影響を及ぼしうるというAIの特性から、特定の分野だけ独立したガバナンスを適用するということは難しい。原子力分野のAIガバナンスも国レベルのガバナンスと密接な関係を持ち、整合を取る必要がある。そのため、各国の原子力分野でのAIガバナンスを理解するためには、国（あるいはEU）レベルでのガバナンスを理解する必要がある。そこで本章ではAIガバナンスの概要を示すとともに、世界レベル、国レベルのAIガバナンスの概要を紹介する。

6.1 AIガバナンスの構造

本稿作成時の状況では、国連、OECDなどが、強制力を持たない国際的なガイドラインを制定している。各国政府やEUはこの国際的なガイドラインと整合する形で国レベルのガイドライン、法律を制定している。国レベルのガバナンスは、EUのAI法のように強制力を伴う法律から、強制力を伴わないガイドラインまで様々である。

現時点では、AI法が全分野に水平的に適用されるEUでは、分野固有の問題は標準（規格）で吸収するようである。また加盟各国が所轄当局を持ち、実際の施行は各国政府（原子力分野の場合は各国の規制当局）が担当すると見られる。一方、英国はできるだけ政府として新たな規制法を制定せず、現状の法律の枠組みの中で各規制当局がガイドライン等を制定する方針である。しかし、規制当局単独で複雑なAIモデルの安全性検証を行うのは現実的ではなく、AIの安全性を研究AIセーフティー・インスティテュートや、民間の検証機関等が個別のAIモデルの安全性を検証するようである。日本も英国と近い規制方針で、各分野のAIガバナンスは現在の規制当局がガイドライン等を整備することになると見られる（日本政府の今後の方針については6.6.6を参照）。

AIシステムは従来のシステムとは大きく異なるため、既存の規制制度をうまく利用することは難しいと考えられている。そのため、本章で紹介する例では、現実からは隔離されるか、あるいは現実世界ではあるが制御の容易な「サンドボックス」と呼ばれる環境でAIシステムの挙動を確認するサンドボックス制度により審査を行っているか、行う予定の国が多い。

6.2 国際的なAIガバナンス

OECDの人工知能に関する理事会勧告（Recommendation of the Council on Artificial Intelligence）[6-1]は2019年5月という比較的早い時期に採択され、多くの国のガイドライン等で参照されている。このガイドラインでは、

AI システムとは、人間が定義した一定の目的のために、実環境あるいは仮想環境に影響を及ぼす予測、推薦又は意思決定を行う機械ベースのシステムである。AI システムは様々なレベルの自律性を備えて稼働するよう設計されている。

と定義しており、多くの国のガイドラインがこの定義を踏襲している（日本語訳は総務省の非公式翻訳[6-2]による）。また、以下の 5 つの原則を推進かつ履行するよう勧告しており、各国のガイドラインもこれらの原則を踏襲している。

1. 包括的な成長、持続可能な開発及び幸福
   人間の能力の増強や創造性の向上、少数派の包摂の促進、格差の改善、及び自然環境の保護などがもたらす包摂的な成長、持続可能な開発及び幸福の増進といった人々と地球にとって有益な結果を追求すること。
2. 人間中心の価値観及び公平性
   法の支配、人権及び民主主義の価値観を尊重するために、人間による最終的な意思決定の余地を残しておくことなど、状況に適した形で、かつ技術の水準を踏まえたメカニズムとセーフガードを実装すること。
3. 透明性及び説明可能性
   AI システムに関する透明性と責任ある開示に積極的に関与するために、a) AI システムの一般的な理解を深めること、b) 職場におけるものを含め、AI システムが使われていることをステークホルダーに認識してもらうこと、c) AI システムに影響される者がそれから生じた結果を理解できるようにすること、及び d) AI システムから悪影響を受けた者がそれによって生じた結果に対して、その要因に関する明快かつ分かりやすい情報、並びに予測、推薦又は意思決定のベースとして働いたロジックに基づいて、反論することができるようにすること。
4. 頑健性、セキュリティ及び安全性
   AI システムは、通常の使用、予見可能な使用や誤用、又はその他の悪条件においても 正常に機能するとともに、不合理な安全上のリスクをもたらすことがないように、そのライフサイクル全体にわたって頑健で、セキュリティが高く、かつ安全であるべきである。
5. アカウンタビリティ
   AI のアクターは、その役割と状況に基づき、かつ技術の水準を踏まえた形で、AI システムが適正に機能していること及び上記の原則を尊重していることについて、アカウンタビリティを果たすべきである。

さらに幅広い国際的合意としては、2024年3月に国連総会で「持続可能な開発のために安全で安心、信頼できる人工知能システムの機会をつかむ」という決議が採択された[6-3, 6-4]。

この決議は米国が主導し、日本も含む多数の国が共同提案しており、今後国際的な規範になると考えられる。

近年飛躍的な性能向上を遂げている、大規模言語モデルをはじめとした生成 AI（あるいは基盤モデル）についてはその高い能力への期待と危機感から、各国で 2023 年に規制の機運が高まった。国際的な枠組みでも G7 が「広島 AI プロセス包括的政策枠組み」[6-5, 6-6] の中で「全ての AI 関係者向けの広島プロセス国際指針」[6-7]、「高度な AI システムを開発する組織向けの広島プロセス国際指針」[6-8]、「高度な AI システムを開発する組織向けの広島プロセス国際行動規範」[6-9] という指針、規範を制定している。このような取組みもあり、高度な AI システムに対する規制は国際的に整合しつつある。

これらの国際的な規範には強制力は無いが、各国や EU の国内的な法律、ガイドライン等は概ね国際的な規範と整合するように作成されている。本節では、その中でも特に原子力分野で特に参考になると考えられる、EU、英国及び米国の AI ガバナンスの概略を述べる。また日本の原子力分野での AI ガバナンスを考える上で不可欠な日本の AI ガバナンスの概略も記述する。

6.3 EU の AI ガバナンス

EU は過去にガイドライン等 [6-10] によるソフトロー的な AI ガバナンスを志向していたが、現在では AI Act（AI 法）によるハードロー規制に舵を切っているため、本節は AI 法を中心に記述する。

EU の AI 法は 2023 年 12 月に合意が成立し [6-11]、2024 年 3 月の欧州議会での可決 [6-12]、5 月の EU 理事会での可決 [6-13] を経て成立が確実になった。今後、官報への掲載等の手続を経て正式に発効し、2026 年中にも完全適用される見通しである [6-14]。本ノート執筆時点では官報掲載版は入手できていないため、AI 法の条文については 2024 年 4 月 19 日付の最終ドラフト [6-15] を参照し、適宜欧州議会によるブリーフィング [6-16] も参考にする。

6.3.1 適用範囲

AI 法は EU 域内、あるいは第三国に拠点を置く AI システムの提供者が、EU 市場に AI システムを投入する場合、または EU 域内で AI システムを使用する場合に適用される。第三国に所在する AI システムの提供者とユーザーにも、そのシステムの生成した出力が EU 域内で使用される場合には適用される。ただし軍事、防衛、又は国家安全保障の目的のみに開発又は使用される AI システム、第三国の公的機関、国際機関、法執行及び司法協力に関する国際協定の枠組みで AI システムを使用する当局には適用されない。

AI 法はほとんどすべての分野に包括的に適用される、水平的（horizontal）な規制であり、EU 加盟国に同一の基準が適用される。原子力分野は除外規定に含まれていないため、AI 法による規制が適用されるものと考えられる。

6.3.2 AI の定義

AI 法では AI システムの提供、使用に法的規制が加えられるため、明確な定義が要求される。AI 法の第 3 条第 1 項（以降、本節で特に明記せず条項を示した場合は AI 法の条項とする）では

> 「AI システム」とは、様々なレベルの自律性で動作するように設計され、配備後に適応性を示す可能性のある機械ベースのシステムであり、明示的または暗黙的な目的のために、物理的または仮想的な環境に影響を与えることができる予測、コンテンツ、推奨、または決定などの出力を生成する方法を、受け取った入力から推測するものである。

と記述している。これは OECD のガイドラインに比較的近いものであり、「機械」の内容は明示していないため、機械学習や深層学習だけでなく、エキスパートシステムなども含まれると考えられる。

6.3.3 リスクベースアプローチ

AI 法の特徴は、禁止まで含んだ「リスクベース」アプローチによるリスクに応じた規制である。CEIMIA（Centre d'expertise international de Montréal en intelligence artificielle）のホワイトペーパー [6-17] によると、リスクに応じて規制のレベルが変わることに関しては英米などと同様であるが、英米が実際に AI システムで引き起こされるであろう「インパクト」を重視しているのに対し、AI 法では AI の種類、使用法による「仮定の」リスクに基づいている。ドラフトの段階ではリスクレベルに応じて、許容できない（Unacceptable）リスク、ハイリスク、限定的（Limited）リスク、低い、あるいは最小限（Low or minimal）リスクの 4 分類だったが、最終的にはシステミック・リスクを持つ汎用目的 AI（第 51~56 条）が追加された。以下では「許容できない AI」、「ハイリスク AI」、「システミック・リスクを持つ汎用目的 AI」について概説する。

(1) 許容できない AI

許容できない AI としては

- 「サブリミナル・テクニック」を展開する AI システム
- 特定の社会的弱者（身体的または社会的障害）を搾取する AI システム
- 公的機関、またはその代理で社会的スコアリング目的で使用される AI システム
- 公的にアクセス可能な空間における、法執行を目的とした「リアルタイム」遠隔生体認証システム（いくつかの例外を除く）

の 4 点が挙げられている。

(2) ハイリスク AI

人々の安全や EU 基本憲章で保護されている基本的人権に悪影響を及ぼす可能性のある限られた AI システムはハイリスクであるとみなされる。AI 法の付属書にハイリスク AI システムのリストが提示されており、適宜見直されるとされている。

付属書Ⅲ（Annex Ⅲ）で挙げられているハイリスク AI は

- 関連する EU 法又は国内法で認められている生体認証
- 重要なデジタルインフラ、道路交通、水、ガス、暖房、電気の供給における管理・運用の安全コンポーネントとして使用されることを意図した AI システム
- 教育、職業訓練（例えば学習成果の評価、学習の指導、不正行為の監視など）
- 雇用、労働者管理、自営業へのアクセス（例えばターゲット求人広告の掲載、求人への応募の分析・選別、候補者の評価など）
- 必要不可欠な民間・公的サービスや給付（医療など）へのアクセス、自然人の信用度評価、生命保険や医療保険に関するリスク評価と価格設定
- 関連する EU 法又は国内法で認められている、移民、庇護、国境の管理、法執行、国境管理、司法管理、民主主義プロセスの分野で使用される特定のシステム

である。また、自然人のプロファイリングを行う AI システムは常にハイリスク AI とみなされる（第 6 条第 3 項）。一方、付属書のリストに含まれていても、

- 狭い手続き型タスク（注 24）を意図した AI システム
- すでに完了している人間の活動の結果を改善することを目的とした AI システム
- 意思決定パターンの検出や、以前の意思決定パターンからの逸脱を検出することを意図しており、すでに完了している人間の評価に取って代わったり影響を与えたりすることを意図していない AI システム
- 付属書IIIに記載された用途の目的に関連する審査の準備作業を行うことを意図している AI システム

はハイリスク AI とはみなされない。また、第 6 条第 5 項で

> EC は欧州人工知能委員会（後述）と協議の上、本規則（AI 法）の発効後 18 カ月以内に、第 96 条（注 25）に沿って、AI システムにおけるハイリスク及び非ハイリスクの用途の包括的な実例リストとともに、本条を実際に実施するためのガイドラインを提供する。

（注 24） AI 法の前文第 53 項では、「非構造化データを構造化データに変換する AI システム」、「受信文書をカテゴリーに分類する AI システム」、「多数のアプリケーションの重複を検出する AI システム」をいう。狭く限定された性質のタスクを例として示している。

（注 25） 「本規則の実施に関する EC のガイドライン」という条項。

としており、今後ハイリスク AI の具体例が提供されるはずである。ハイリスク AI に課される要求は多岐にわたるが、欧州議会[6-16]によると

- 事前適合性評価の要求
  - ハイリスク AI システムの提供者は、そのシステムを市場に出したり、サービスを開始したりする前に、EC が管理する EU 全体のデータベースに登録することが義務付けられる。
  - 既存の製品安全法制が適用される AI 製品やサービスは、すでに適用されている第三者適合性評価の枠組みに従う。
  - 現在 EU 法が適用されていない AI システムの提供者は、新たな要件に適合していることを示す適合性評価（自己評価）を実施し、CE マーキング（注 26）を使用できるようにしなければならない。
  - バイオメトリクス認証に使用されるハイリスク AI システムのみが認証機関（notified body）による適合性評価を必要とする。
- その他の要求
  - ハイリスク AI システムはリスク管理、テスト、技術的堅牢性、データ訓練及びデータガバナンス、透明性、人による監督、サイバーセキュリティに関する様々な要件に準拠しなければならない。
  - ハイリスク AI システムの提供者、輸入業者、販売業者、ユーザーはこれらの義務を果たさなければならない。
  - EU 域外の提供者は、EU 域内に認可された代理店を置き、適合性評価を確実に行い、市販後のモニタリングシステムを確立し、必要に応じて是正措置を講じる必要がある。
  - 現在策定中の、（AI 法に）整合した EU 規格（harmonized EU standards）に適合する AI システムは、AI 法への適合が推定される。

とある。Veale ら[6-18]によると、多くの製品では AI 法に整合した規格（標準）に準拠することで AI 法を満たすと見られるが、原子力分野には AI 法に整合した規格が存在しないため、前述の「適合性評価（自己評価）」が適用されると考えられる。

(3) システミック・リスクを持つ汎用目的 AI

最終合意案で新たに加わった分類である。第 3 条第 63 項で

> 「汎用目的 AI モデル（general purpose AI model）」とは、大規模な自己教師あり学習を使用して大量のデータで訓練された場合を含め、モデルが市場に投入される方法にかかわらず、重要な汎用性を示し、広範囲の明確なタスクを適切に実行す

（注 26） 製品が EU 基準に適合していることを示すマーク。

る能力を有し、様々な下流のシステムやアプリケーションに統合できる AI モデルを意味する。これには、研究、開発、プロトタイピング目的で、市場にリリースされる前に使用される AI モデルは含まない。

と定義されている「汎用目的 AI モデル」は、一般的に基盤モデル（foundation model）などと呼ばれているものに相当する。また、第 3 条第 65 項で

> 「システミック・リスク（systemic risk）」とは、汎用目的 AI モデルの影響力の大きい能力に特有のリスクであり、その影響力の大きさにより（欧州）連合市場に重大な影響を及ぼすリスク、または公衆衛生、安全、治安、基本的権利、社会全体に対する実際、もしくは合理的に予見可能な悪影響により、バリューチェーン[注 27]全体に大規模に伝搬しうるリスクを意味する。

と「システミック・リスク」について定義している。第 51 条では「システミック・リスクを持つ汎用目的 AI」について記述しており、第 2 項で訓練時の計算量が10^{25} FLOPs（FLoating-point Operations）を超えるものという数値基準が提示されている。Webb[6-19]は OpenAI 社の最新の大規模言語モデル GPT-4 の訓練時の計算量を2×10^{25} FLOPs 程度と推定しており、現時点での最先端の汎用目的 AI モデルがこの分類に該当すると見られる。また Google など[6-20]によると、大規模言語モデルの性能は訓練時の計算量が10^{23}~10^{24} FLOPs を超えると「創発的能力（emergent ability）」を獲得して飛躍的に向上することがあるとしており、10^{25} FLOPs という閾値には一定の妥当性がある。

システミック・リスクを持つ汎用目的 AI モデルの提供者には、リスクの評価と軽減、重大インシデントの報告、最先端のテストとモデル評価の実施、サイバーセキュリティの確保、モデルのエネルギー消費に関する情報の提供義務などの要求が記されている（第 52 条）。

6.3.4 サンドボックス

限定的な環境で AI システムの開発・検証を行う環境を提供するサンドボックス制度については 3.3.3 で述べた通りである。

6.3.5 罰則

AI 法の罰則については 3.2.5 で述べた通りである。

6.3.6 AI 法の実施

AI 法の適用と実施を監督し、市場監視活動を実施するため EU 加盟各国は 1 つ以上の国家所轄当局を指定する。効率性を高め、国民やその他の関係者との公式な窓口を設置する

[注 27] ここでは EU 市場全体の活動を示していると考えられる。

ため、加盟各国は国家監督当局を指定し、国家監督当局は欧州人工知能理事会（European Artificial Intelligence Board）において国を代表する。また産業界、新興企業、中小企業、市民社会、学会などの利害関係者をバランスよく代表する諮問委員会が技術的な専門知識を提供する。さらに EC は委員会内に新たな欧州 AI 事務局（European AI Office）を設立し、欧州人工知能委員会（European Artificial Intelligence Board）と協力し、独立した専門家からなる科学委員会の支援を受けて汎用目的 AI を監督する。なお、欧州 AI 事務局は 2024 年 5 月に発足した [6-21]。

6.3.7 適用のスケジュール

AI 法は、官報に掲載された翌日から 20 日目に発効し、発効から 24 カ月後に全面適用される（前文 179 項）。また同項には

- 届出機関及びガバナンス構造に関する規定は、発効後 2 カ月後から適用されるべきである。
- 汎用目的 AI モデルの提供者に対する義務は、発効後 12 カ月後から適用されるべきである。
- 行動規範は発効後 9 カ月までに準備されるべきである。
- 罰則に関する規定は、発効後 12 カ月後から適用されるべきである。

と記載されている。重要インフラで使用されるハイリスク AI システムは、発効から 24 カ月後（2026 年）に規制の対象となるとみられる。

6.4 英国の AI ガバナンス

Robertsら[6-22]は英国を、国レベルのAI政策の発表数が米国に次いで2 位であること[6-23]、2016年から2021年にかけての立法文書におけるAIへの言及数が世界一多いこと[6-24]を理由に、AIの適切なガバナンスに強い関心を持っている国であると述べている。その英国はG7の中では最もソフトロー(注28)的なガバナンスを志向しており、当面の間AIの法的な規制は行わないと2023年3月に発行したホワイトペーパー「AI規制へのイノベーション促進アプローチ」[6-25]で表明した。本節では「AI規制へのイノベーション促進アプローチ」を中心に英国のAIガバナンスについて記述する。

6.4.1 AI 規制へのイノベーション促進アプローチ（2023）

現時点での英国政府の AI 規制の方針を示している「AI 規制へのイノベーション促進アプローチ」について概説する。

(注 28) 法的拘束が全く無いか、緩いこと。対義語としては「ハードロー」。

(1) AI の定義

「広くコンセンサスを得られる AI の一般的な定義は存在しない」としながらも、「適応性」と「自律性」という 2 つの特性を参照して定義を行い、この特性により特別の規制対応が必要であるとしている。引用すると、

- AI の「適応性」は、システムの結果の意図や論理を説明することを困難にする可能性がある。
  a. AI システムは、一度又は継続的に「訓練」され、多くの場合、人間には容易に識別できないデータのパターンやつながりを推論することによって作動する。
  b. このような訓練を通じて、AI システムはしばしば、人間のプログラマーが直接想定していない新しい形の推論を実行する能力を開発する。
  c. AI の「自律性」は、結果に対する責任の所在を明確にすることを困難にする。
  d. AI システムの中には、人間の明示的な意図や継続的なコントロールなしに意思決定を行うものもある。

また

> 適応性と自律性の組み合わせは、AI システムの出力や、それらが生成される基礎となる論理を説明、予測、制御することを困難にする。また、システムの運用や出力に対する責任を割り当てることも困難となる。

と述べている。

(2) 5 原則

ホワイトペーパーの Executive summary では経済のあらゆる分野における AI の責任ある開発と利用の指針として、第 3.2.3 節で

- 安全性、セキュリティ、頑健性
- 適切な透明性及び説明可能性
- 公平性
- アカウンタビリティとガバナンス
- 競争可能性と救済

という 5 つの原則を掲げている。この原則は OECD の AI 原則[6-1]を基礎として、英国の取組を反映したものだと述べられている。この原則の実装に関しては

> 我々は最初は、これらの原則の法制化は行わない。企業に対して新たに厳格で面倒な法的要件を課すことは、AI のイノベーションを抑制し、将来の技術進歩に迅速かつ適切に対応する能力を低下させる恐れがある。その代わりに、この原則は非法

定ベースで発令され、既存の規制当局によって実施される。このアプローチは、規制当局の固有分野の専門知識を利用して、AI が使用される特定の状況に合わせて原則の実施を調整する。導入の初期期間中、我々は規制当局と協力し、原則の適切な適用を妨げる障壁を特定し、法によらない規制の枠組みが望ましい効果を上げているかどうかを評価する。

と述べている。また規制そのものは法制化しないものの

導入期間の後、国会の時間が許すならば、規制当局に対してこの原則を十分に考慮することを求める法的義務を導入する予定である。規制当局、産業界、学識経験者からは、この枠組みの実施を支援するためのさらなる措置を導入すべきとの意見もあった。規制当局に原則への配慮を求める義務を課すことで、規制当局が特定の状況において原則を適用する場合の判断を柔軟に行えるようにする一方、原則を実施する義務を強化することができるはずである。枠組みをモニターした結果、規制当局と協力し、適応可能なアプローチをとるという我々の提案に沿って、法制化せずに実施することが効果的であることが分かれば、そのような法制化は行わない。

とあり、原則を考慮した規制の導入を規制当局に求めることの法制化には含みを残している。また

我々は、全体的な枠組みが規制全般にわたるイノベーションを促進しつつ、リスクに対して適切かつ効果的な対応を提供するために必要な、中心的なサポート機能をいくつか特定した。

として

- 全体的な規制の枠組みの有効性と原則の実施（実施がどの程度イノベーションを支援しているかも含む）のモニタリングと評価。これによって私たちは敏感であり続け、AI の能力や技術の進展の中で効果的であり続けるために適応が求められる場合も含め、必要であれば枠組みを適応させる。
- AI から生じる経済全体のリスクを評価・監視する。
- AI 技術の新たなトレンドへの首尾一貫した対応を知らしめるため、産業界の招集も含め、ホライゾン・スキャニング（将来展望活動）とギャップ分析を実施する。
- AI イノベーターが新技術を市場に投入できるよう、テストベッドやサンドボックス構想を支援する。
- 枠組みの継続的な改善の一環として、教育と関心を提供して企業に透明性を与え、市民が声を上げることができるようにする。
- 国際的な規制の枠組みと相互運用性を促進する。

という政府からの支援も記載している。ただし、

> 上記の活動は、規制当局が行っている業務に取って代わるものでも重複するものでもなく、新たな AI 規制機関の設立を伴うものでもない。

と、あくまでも既存の規制機関が中心的な役割を果たすことが明言されている。

(3) 状況特化

第 3.2.2 節で、EU のリスクベースアプローチに対して、英国の枠組みは状況特化（context-specific）だと述べている。広い意味ではリスクベースではあるが、

> セクターやテクノロジー全体にルールやリスクレベルを割り当てることはしない。その代わりに、特定の用途において AI が生み出しそうな結果に基づいて規制を行う。例えば、重要インフラにおける AI の全ての用途をハイリスクに分類することは、つり合いが取れず、効果的でもないだろう。重要インフラにおける AI の用途の中には、機械の表面的な傷の識別のように、比較的リスクが低いものもある。同様に、オンライン衣料品小売業者の顧客サービス要求のトリアージ（優先順位付け）に使用される AI 搭載チャットボットは、医療診断プロセスの一部として使用される同様のアプリケーションと同じように規制されるべきではない。

とあり、CEIMIA のホワイトペーパー [6-17] が分類するところの「インパクトベース」アプローチになっている。また状況特化によるリスク評価には、単に AI の使用によって生じるリスクだけではなく

> 規制当局は、AI のリスク評価には AI の能力を利用できなかった場合も含めるべきだと我々に語った。例えば、重工業から個人のヘルスケアに至るまで、安全性が重要な業務において AI を利用できないことには、大きな機会費用が発生する可能性がある。状況に敏感になり、枠組みがリスクレベルに比例した対応をすることで、イノベーションを阻害したり、AI がもたらす社会的便益を利用する機会を逃したりすることを避けられる。

と、AI を使わないことによる機会費用の損失も考慮に入れて規制を行うべきだと述べている。

(4) 個々の規制当局

第 3.2.5 節で、「新たな規制の枠組みを実施するにあたり、規制当局には以下を期待する」とし、規制当局の役割として

- 分野横断的な原則を評価し、自らの権限に属する AI の用途に適用する。

- 産業界が原則を適用することを支援するため、原則と既存の法律との相互作用に関するガイダンスを発行する。そのようなガイダンスはまた、コンプライアンスがどのようなものかを説明する。
- それが適切な場合には共同ガイダンスを含む、明確で一貫性のあるガイダンスを共同で作成することにより、複数の規制当局の権限内で事業を行う企業を支援する。

と述べている。また、第 3.2.6 節では「原則の適用に関する規制当局へのガイダンス」として

> 規制当局及び産業界との取組みでは、中央政府が規制当局を支援する必要性が浮き彫りになった。私たちは規制当局と協力し、枠組みがどのように運用されるべきかという私たちの期待に沿った形で原則を実施するのに役立つガイダンスを作成する。既存の法的枠組みは、すでに規制当局の行動を義務付け、指導している。例えば、ほぼ全ての規制当局は「規制当局規範」に拘束され、公的機関として人権法を遵守することが求められている。私たちが提案する規制当局へのガイダンスは、原則を適用する際に、規制当局が以下のように支援され、奨励されることを保証するものである:
>
> ① 特定の状況において AI がもたらすリスクに焦点を当てることで、成長とイノベーションを促進するつり合いの取れたアプローチを採用する。
> ② 政府が実施、あるいは政府に代わって実施する分野横断的なリスク評価を考慮し、優先されるリスクに対処するためのつり合いの取れた措置を検討する。
> ③ 適切な規制要件を策定、実施、執行し、可能であれば、原則の実施を既存の監視、調査、執行のプロセスに統合する。
> ④ 必要に応じて、原則と関連する規制要件への業界のコンプライアンスを支援するための共同ガイダンスを策定する。
> ⑤ 保証技術や技術標準のような、信頼できる AI のためのツールが、どのように規制遵守を支援できるか検討する。
> ⑥ 政府による枠組みの監視と評価に積極的かつ協力的に関与する。

と述べている。なお、この「ガイダンス」は（「AI 規制へのイノベーション促進アプローチ」の）出版後 6 カ月以内に発行されることになっている。

(5) サンドボックス

Roberts ら [6-22] によると英国は規制のサンドボックス使用に関してはパイオニアであり、「AI 規制へのイノベーション促進アプローチ」でも AI サンドボックスについて第 3.3.4 節で

規制のサンドボックスとテストベッドは、我々が提案する規制体制において重要な役割を果たすだろう。このような取組みによって、政府と規制当局は以下のことが可能になる:

- イノベーターが斬新な製品やサービスをより早く市場に投入し、経済的・社会的利益を生み出せるよう支援する。
- 規制の枠組みが実際にどのように運用されているかを検証して、対処すべきイノベーションに対する不必要な障壁を明らかにする。
- 規制の枠組みが適応する必要のある、新たな技術や市場のトレンドを特定する。

と述べている。

(6) 規制当局の能力

第 3.3.5 節では、「規制当局の能力」として

現在のところ、規制当局の権限を拡大することは考えていないが、AI の利用を効果的に規制するためには、多くの規制当局が新たなスキルや専門知識を身につける必要がある。我々の調査では、AI を理解し、そのユニークな特徴に対処する能力に関して、規制当局官にレベル差があることが浮き彫りになった。また AI リスクに対処するために規制当局が必要とする能力や、規制当局がそれらを習得するための最善の方法についても、幅広い意見が得られた。

と述べ、特に、以下の専門知識について AI に関する能力格差があるとしている。

- AI の技術に関する専門知識。例えば、製品やサービスを提供するために AI がどのように使用されているか、技術標準の開発、使用、適用可能性について。
- AI のユースケースが複数の規制制度間でどのように相互作用するかについての専門知識。
- AI 技術が既存のビジネスモデルを破壊するためにどのように利用されているか、潜在的な機会と規制目標に影響を与えうるリスクの両方に関するマーケットインテリジェンス。

また、規制当局の能力としては

- AI のユースケースやアプリケーションの出現に効果的に適応し、組織全体でこの知識を吸収、共有する。
- 保証技術を提供する組織（保証サービスプロバイダーなど）や技術標準を開発する組織（標準開発組織など）と連携し、関連するツールを特定し、規制の枠組みやベストプラクティスに組み込む。
- 複数の規制制度にまたがる AI のユースケースの規制において、規制当局間で知

識を共有し、協力する。
- 通常は管轄外の組織や団体との関係を構築し、効果的にコミュニケーションを図る。

を挙げている。

(7) 信頼できる AI のためのツール

第 4 部（Part 4）では信頼できる AI のための 2 つのツールを紹介している。

① AI 保証技術（AI assurance）

CEIMIA のホワイトペーパー [6-17] によると、英国の AI 規制は「ライトタッチ」であるため、自己監視と執行に大きな役割があり、第三者監査（third-party audit）の役割が大きくなる。「AI 規制へのイノベーション促進アプローチ」によれば AI システムの効果的保証には「影響評価、監査、性能テスト、形式的検証手法などが含まれる」。英国の CDEI（Centre for Data Ethics and Innovation）が 2021 年に発行した AI 保証のロードマップ [6-26] では

> 私たちのビジョンは、英国が今後 5 年以内に、繁栄し、効果的な AI 保証エコシステムを構築することです。革新的な新興企業やスケールアップ企業とともに、強力な既存のプロフェッショナル・サービス企業が、AI における正当な信頼を構築するための様々なサービスを提供するでしょう。

と、民間企業による AI 保証産業の育成を目指している。また、「AI 規制へのイノベーション促進アプローチ」でも AI 保証について触れていて、2023 年春に AI 保証テクニックのポートフォリオを発行するとしている（6.4.3 で説明する「AI 保証入門」）。

② AI 技術標準（AI technical standards）

「保証技術は技術標準によって裏打ちされる必要がある」とし、英国が国際的な技術標準の策定において主導的な役割を果たしていると述べている。

(8) 今後の展開

今後の展開について主なものをまとめると（期限は「AI 規制へのイノベーション促進アプローチ」の出版時点である 2023 年 3 月が起点になっていると考えられる）、

① 6 カ月以内
- 公共部門、規制当局、学会、市民社会などと協議を行い、それに対する回答を公表する。
- 規制当局が枠組みを実施するための分野横断的原則と初期ガイドラインを発

表する。

- 政府が枠組みの中心的機能を果たす AI 規制ロードマップを発表し、AI サンドボックスやテストベッドを試験的に導入する。
- 委託研究プロジェクトから得られた知見を分析し、潜在的なコンプライアンス障壁、AI ライフサイクルの説明責任、規制当局の枠組みの実施能力とそれへの支援、AI リスク測定と報告のベストプラクティスなどの理解を深める。

② 6~12 カ月

- 枠組みの中心的機能を実現するためのイニシアティブとパートナーシップを確立する。
- 主要な規制当局に対し、分野横断的な原則がその権限内でどのように適用されるかについてのガイダンスを公表するよう促す。
- 中央監視・評価機能の在り方に関するアイディアを提案し、これらの提案を利害関係者の協議に供する。
- イノベーターや規制当局と規制のサンドボックスの開発を継続する。

③ 1 年後以降

- 枠組みを効果的にするための中心的機能を提供する。
- ガイダンスを公表していない規制者にガイダンスを公表するよう求め、支援する。
- パイロット版から得られた知見に基づいて、規制のサンドボックスやテストベッドを開発する。
- 最初のモニタリング・評価報告書を公表する。
- AI 規制ロードマップを更新する。

となる。すでに「AI 規制へのイノベーション促進アプローチ」の発行から 1 年近く経過しているが、必ずしも 12 カ月後が期限の全ての取組みの結果が公表されているわけではない。今後、取組みの進行状況が公表されていくものと考えられる。

6.4.2 AI セーフティー・インスティテュート

英国では先進的な AI を「フロンティア AI」（frontier AI）としているが、EU の「システミック・リスクを持つ汎用目的 AI」に近い概念であると考えられる。フロンティア AI の安全で信頼できる発展を目的として、英国政府はフロンティア AI タスクフォース（Frontier AI Taskforce）を 2023 年 4 月に設立したが、先進的な AI の急激な進化に合わせて 2023 年 11 月にこの組織を AI セーフティー・インスティテュート（AI Safety Institute: AISI）として改組した[6-27]。AISI の中核的機能としては

- 先進的な AI システムの開発と評価。

- さまざまな探索的研究プロジェクトを立ち上げて AI の安全性に関する基礎研究を推進する。
- 政策立案者、国際的パートナー、民間企業、学術界、市民社会、一般市民など、国内外の関係者との情報交換を促進する。

ということが挙げられており、英国の AI ガバナンスで重要な役割を果たすと考えられる。ただし AISI が規制機関でなく、政府の規制を決定することがないということは明確に述べられている。また、今後日米の AI セーフティー・インスティテュートと協力を進めていくものと見られる。

6.4.3 AI 保証

「AI 規制へのイノベーション促進アプローチ」に述べられていたように、第三者機関による AI 保証は英国の AI ガバナンスの重要な柱である。英国はサイバーセキュリティ産業のように、民間企業を中心とした AI 保証エコシステムの構築を目指しているようであり、その第一歩として 2024 年 2 月に「AI 保証入門」[6-28] を発行している。「AI 保証入門」の 3.3 節では、英国の AI 規制方針が

> AI がもたらす潜在的なリスクは、そのアプリケーションの状況によって異なることを認識したうえで、特定の状況において AI がもたらす特有の課題と機会のため、英国の AI ガバナンスへのアプローチは技術そのものではなく、結果に焦点を当てている。この結果ベースのアプローチを実現するため、既存の規制当局は、各分野における規制原則の解釈と実施に責任を負い、各々の分野内でこれらの結果を達成する方法に関する明確なガイドラインを確立する。組織が説明責任を負うことができる検証可能な主張を行い、評価するプロセスの輪郭を描くことで、AI 保証は、より広範な AI ガバナンスと規制の重要な側面となる。

と端的に示されている。また、英国政府は 2023 年 6 月に AI 保証技術の例を紹介する「AI 保証技術のポートフォリオ」[6-29] を発行している。その中で

> 英国認証機関認定審議会（United Kingdom Accreditation Service: UKAS）は英国唯一の国家認定機関である。UKAS は、AI 保証を含むさまざまなサービスを提供する適合性評価機関として知られる第三者機関を評価するために、政府によって任命される。

とあり、UKAS の認定を受けた第三者機関が AI 保証を担うと見られる。

6.4.4 アラン・チューリング研究所

アラン・チューリング研究所（Alan Turing Institute）は 2015 年に設立されたデータサイ

エンスの国立研究所で、2017 年に政府の勧告により AI も任務に含めている [6-30]。AI 標準ハブを主導するなど、英国の AI ガバナンスで重要な役割を果たしている。

(1) AI 標準ハブ

AI 標準ハブ（AI Standards Hub）[6-31] は英国規格協会（British Standards Institution: BSI）、イギリス国立物理学研究所（National Physical Laboratory: NPL）との連携でアラン・チューリング研究所が主導しているイニシアティブである。標準が AI ガバナンスツール及びイノベーションメカニズムとして果たしうる役割に焦点を当て、信頼でき、責任ある AI を推進することを使命とし、

- AI の標準化に関する議論を形成し、健全で一貫性のある効果的な標準の開発を促進する。
- 国内及び国際的な AI ガバナンスの実践の情報提供、強化を行う。
- AI 標準策定へのマルチステークホルダーの参加を促進する。
- 公表された関連企画の評価と利用を促進する。

という目標を示している。

6.4.5 将来の法規制の可能性

これまで述べたように、英国政府は現時点では AI に対して新たな法規制を行わない方針であるが、2024 年 4 月に、英国政府が大規模言語モデルへの規制を検討しているという報道 [6-32] がなされた。

6.5 米国の AI ガバナンス

米国の AI ガバナンスは、各規制機関がそれぞれの分野ごとにソフトロー的な規制を行うという英国に近いスタイルになっている。英国があくまでも新規な法制化を行わず、ホワイトペーパーにより各規制機関に非法的な規制の作成を求めていたのに対し、米国では AI に関する大統領令、連邦法が制定されている。しかし、米国の AI ガバナンスに関する大統領令、連邦法の多くは政府各機関に AI 規制の作成を求めたり、政府機関内での AI 使用の振興やルール策定、あるいは AI に対応できる人材を増やすことを求めたりなど、政府機関への対応を求めるものであり、直接民間の AI 開発や使用を規制する内容は少ない。現状では米国内での民間の AI の使用、開発に関するルールは法的拘束力を持たない「AI 権利章典の青写真」[6-33] や NIST の「AI リスクマネジメントフレームワーク」[6-34, 6-35] が中心となる。また EU や英国とは異なり、米国の AI ガバナンスに関する大統領令、連邦法は体系的に整理されているわけではない。本報告では米国の AI ガバナンスを、大統領令、連邦法ごとに NRC に関連が大きそうな項目を中心として内容を紹介する。

6.5.1　AIにおけるアメリカのリーダーシップの維持（2019）

トランプ政権は2019年2月に大統領令「AIにおけるアメリカのリーダーシップの維持（Maintaining American Leadership in Artificial Intelligence）」（EO 13859）[6-36]を発令した。米国議会調査局（CRS）のレポート[6-37]によれば、この大統領令ではAIの研究開発の投資と調整、連邦政府のデータ、モデル、計算資源をAI開発に利用できるようにすること、AI技術使用への障壁を減らすこと、AIイノベーションに関する技術的・国際的基準を策定すること、AIと国家安全保障上の懸念に関する行動計画を作成すること、AIを開発・利用できる人材を育成することなどが定められている。Section 6は「AIアプリケーションの規制に関するガイダンス」（Guidance for Regulation of AI Application）となっており、AI規制の方針について記述されている。このセクションの(a) には

> 本命令の日付から180日以内に、行政管理予算局（OMB） 長官は、科学技術政策局（OSTP）長官、国内政策会議（Domestic Policy Council）委員長、国家経済会議（National Economic Council）委員長と連携し、またOMB長官が決定するその他の関連機関及び主要な利害関係者と協議の上、全国家機関の長に対し、以下の事項を明記したメモを送付しなければならない。
>
> ① AIによって強化される、あるいは可能になる技術や産業分野に関し、当該機関による規制的・非規制的アプローチの開発に情報を提供し、市民の自由、プライバシー、米国の価値観を守りつつ、米国のイノベーションを促進する。

とあり、OMB長官による政府機関へのメモの送付が指示されている。また各機関がそれぞれ規制・非規制アプローチを開発するとしていることから、EUのような水平的な規制ではなく、英国のような分野別の規制を志向していることが明らかである。同じセクションの(c) には

> 本項(a)に記載されたメモの日付から180日以内に、規制当局を有する機関の長は、AIアプリケーションに関連する権限を見直し、メモとの整合性を達成するための計画をOMBに提出しなければならない。

とOMB長官の覚書に基づいた計画の策定を求めている。さらに、このセクションの(d) では

> 本命令の日付から180日以内に、商務長官（the Secretary of Commerce）は、NIST所長を通じて、AI技術を利用した信頼性、堅牢性、信用性の高いシステムを支援する技術標準と関連ツールの開発に連邦が関与するための計画を発表する。NISTは、商務長官が決定する関係省庁の参加を得て、本計画の策定を主導する。（以下略）

と NIST 主導の標準策定を定めている。

(1) AI アプリケーション規制ガイダンス

「AI におけるアメリカのリーダーシップの維持」で 180 日以内の発出が求められていた OMB 長官のメモ「AI アプリケーション規制ガイダンス」[6-38] は大統領令の発令から 1 年 9 カ月以上経った 2020 年 11 月 17 日に送付された。なお、「政府による AI の利用は本覚書の範囲外である」と明確に述べられている。

「AI アプリケーション規制ガイダンス」では AI アプリケーションの管理原則（Principles for the Stewardship of AI Applications）として以下の 10 の原則が示されている。

① AI に対する社会の信頼（Public Trust in AI）

② 一般参加（Public Participation）

③ 科学的誠実さと情報の質（Scientific Integrity and Information Quality）

④ リスク評価と管理（Risk Assessment and Management）

⑤ 利益とコスト（Benefits and Costs）

⑥ 柔軟性（Flexibility）

⑦ 公平性と無差別（Fairness and Non-Discrimination）

⑧ 情報開示と透明性（Disclosure and Transparency）

⑨ 安全とセキュリティ（Safety and Security）

⑩ 機関間調整（Interagency Coordination）

特に④の「リスク評価と管理」では

> AI に対する規制・非規制のアプローチは、様々な機関や様々な技術にまたがるリスク評価とリスク管理の一貫した適用に基づくべきである。（中略）むしろ、どのリスクが許容可能で、どのリスクが許容できない危害、あるいは期待される利益よりも期待される費用が大きい危害の可能性があるかを判断するために、リスクベースのアプローチを用いるべきである。機関は、リスクの評価について透明性を保ち、説明責任を促進するために、適切な間隔でその仮定と結論を再評価すべきである。これに対応して、AI ツールが失敗した場合、あるいは成功した場合の結果の大きさと性質は、リスクを特定し緩和するために適切な規制努力のレベルと種類を知らせるのに役立つ。（中略）リスクの評価は、そのリスクと、問題となっている AI の適用がなければ得られたであろう状況のリスクとの比較によって行われるべきである。AI の適用によってリスクが軽減されるのであれば、関連する規制は、その適用を許可するのが妥当であろう。

と述べており、AI に対する規制はリスクベース、より詳細には実際に生じると見込まれるインパクトに基づく、CEIMIA のホワイトペーパー [6-17] で言うところの「インパクトベース」アプローチであることを求めている。またリスク評価は AI を使わない場合のリスクと比較することによって評価すべきであるとしている。

AI を導入した場合の便益とリスク評価については⑤「利益とコスト」で

> EO 12866 [注 29] は「正味の利益（潜在的な経済的、環境的、公衆衛生と安全上の利益、その他の利点、分配の影響、及び公平性を含む）を最大化するアプローチを選択する」ことを機関に求めている。各機関は、AI アプリケーションの開発・展開に関連する規制を検討する際、法律に準じて、社会全体のコスト、便益、分配効果を慎重に検討すべきである。このような検討には、AI がそれを補完または代替するために設計されたシステムと比較した場合の、AI を採用することの潜在的な便益とコスト、AI の導入によって生じるエラーの種類が変わるかどうか、他の既存のシステムで許容されるリスクの程度との比較などが含まれる。現行のシステムやプロセスとの比較ができない場合は、システムを導入しない場合のリスクやコストについても評価する必要がある。

と述べ、EO 12866 にしたがって「正味の利益を最大化する」ことが求められている。さらに非規制的アプローチ（Non-Regulatory Approach to AI）として

> ある特定の AI アプリケーションを検討した結果、既存の規制で充分である、あるいは新たな規制のメリットがその時点あるいは予測可能な将来においてそのコストを正当化するものではないと判断する場合がある。このような場合は、当局は、いかなる措置も講じないか、あるいは、その代わりに、特定の AI アプリケーションがもたらすリスクに対処するために適切と思われる非規制的アプローチを特定することを検討することができる。

と述べている。

最後に各機関に対し、2021 年 5 月 17 日までに AI アプリケーションに関連する権限を見直し、「AI アプリケーション規制ガイダンス」との整合性を達成するための計画を提出することを要求している。3.3.4 で取り上げた NRC の AI 戦略計画はこの要求に対応していると見られる。

6.5.2 連邦政府における信頼できる AI の利用促進（2020）

トランプ政権によって 2020 年 12 月に発令された「連邦政府における信頼できる AI の利用促進」（EO13960）[6-39] は、米国議会調査局（CRS）のレポート [6-37] では、「国民の信頼と

（注 29） 「規制の計画と見直し」と題された大統領令（1993）。

信用を醸成するため、連邦政府における AI の設計、開発、取得、使用に関する共通の原則を定め、OMB に対し、各省庁間で原則を実施するための政策指針を策定するよう指示している」、と説明されている。この大統領令の第 3 項は「政府における AI 利用の原則」（Principles of Use of AI in Government）となっており、9 つの原則が掲げられている。ここではその項目だけ紹介する。

① 合法的であり、我が国の価値観を尊重する（Lawful and respectful of our Nation's values）
② 目的意識を持ち、実績を重視する（Purposeful and performance-driven）
③ 正確で、信頼でき、効果的であること（Accurate, reliable, and effective）
④ 安全、安心で、弾力性があること（Safe, secure, and resilient）
⑤ 理解可能であること（Understandable）
⑥ 責任があり、追跡可能であること（Responsible and traceable）
⑦ 定期的な監視（Regularly monitored）
⑧ 透明性が高いこと（Transparent）
⑨ 責任を負うこと（Accountable）

6.5.3 2020 年国家 AI イニシアティブ法（2021）

「2020 年国家 AI イニシアティブ法」[6-40] は連邦議会によって制定された連邦法で、国家 AI イニシアティブの設立などが定められている。CRS のレポート [6-37] に掲載された要約から本報告と特に関係の深そうなものを抜粋すると

- AI 標準の開発をサポートし、信頼できる AI システムのためのリスク管理フレームワークを開発し、AI システムの訓練に使用されるデータセットを文書化して共有するためのベストプラクティスを開発するよう NIST に指示する。（第 5301 項）
- AI ツール、システム、機能、労働力のニーズを進歩させ、DOE の使命に関連する AI 手法とソリューションの信頼性を向上させるための横断的な研究開発プログラムを実行するよう、DOE に指示する。（第 5501 項）

とあり、後述の AI リスクマネジメントフレームワーク [6-33] 作成や、DOE による AI 研究推進の法的根拠になっている。

6.5.4 2020 年政府における AI 法（2020）

同じく連邦法の「2020 年政府における AI 法」[6-41] では、第 4 項「政府機関の AI 利用のための指針」（Guidance for Agency Use of Artificial Intelligence）で OMB 長官に、各政府機関の長に対し AI 利用方針の策定等を求める覚書を発出すること、それを受けて各政府機

関の長には AI 利用計画、または AI を使用せず今後も使用しないという書面による決定をウェブサイトに掲載するよう求めている。

第 5 項「AI の職業シリーズの更新」（Update of Occupational Series for Artificial Intelligence）では OPM 局長に、本法の施行日から 18 カ月以内に、AI に関連する職位に必要なスキルと能力を特定し、既存の職業シリーズを更新及び改善して、AI に関連する主要な職務を含めることを求めている。また現在 AI に関連する役職についている連邦政府職員の数を機関ごとに推定し、各機関が 2 年、5 年後に雇用する必要のある AI に関連する職員数の予測を作成するよう求めている。

6.5.5 AI 権利章典の青写真（2022）

米国政府が 2022 年 10 月に公表した「AI 権利章典の青写真」[6-33] は法的拘束力のないホワイトペーパーだが、内閣府科学技術・イノベーション推進事務局の解説 [6-42] によれば「AI を含む『自動化システム』を構築しガバナンスする際に、米国国民の人権を保護しつつ民主主義的価値を推進するための政策及び実践方法の開発のサポートを目的」としており、米国の AI 原則的な位置づけにあると考えられる。このホワイトペーパーでは

(1) 安全で効果的なシステム（Safe and Effective Systems）

(2) アルゴリズム由来の差別からの保護（Algorithmic Discrimination Protections）

(3) データのプライバシー（Data Privacy）

(4) ユーザーへの通知と説明（Notice and Explanation）

(5) 人による代替手段、配慮、代替システム（Human Alternatives, Consideration, and Fallback）

という 5 つの原則とそれを実践するための具体的な手順と例が記載されている。

6.5.6 安心、安全、信頼できる人工知能の開発と利用（2023）

バイデン政権によって 2023 年 10 月に発令された大統領令「安心、安全、信頼できる AI の開発と利用」（EO14110）[6-43] は AI 技術の潜在的なリスクに対する懸念の高まりを受けたものである。この大統領令は ChatGPT に代表される、大規模なリスクをもたらしうる基盤モデルに大きな焦点を当てているが、米国政府が AI 技術の採用に力を入れることも示している。多くの条項が、連邦政府が AI を安全かつ安心に使用すること、高度な AI ツールや関連産業の国内開発を促進することを奨励している。また、多くの事項について短期間（多くは 120、180 日以内）での実行を求めていることも特徴的である。

(1) 用語の定義

AI については、

人間が定義した所定の目的に対して、現実または仮想環境に影響を与える予測、推奨、または決定を行うことができる機械ベースのシステムである。AI システムは、機械及び人間ベースの入力を使用して現実環境及び仮想環境を認識し、自動化された方法で分析を通じてそのような認識をモデルに抽象化し、モデルの推論を使用して情報や行動の選択肢を策定する。

と定義している。また「デュアルユース基盤モデル」を

広範なデータで学習され、一般的に自己教師あり学習を使用し、少なくとも数百億のパラメーターを含み、広範な文脈に適用可能であり、かつ、安全保障、国家経済安全保障、国家公衆衛生もしくは安全、またはそれらの組み合わせに重大なリスクをもたらすタスクにおいて、以下のような高水準のパフォーマンスを示すか、または示すように容易に修正できる AI モデルを意味する。

① 非専門家が化学、生物、放射線、核（CBRN）兵器を設計、合成、入手、使用するための参入障壁を大幅に下げる。
② 広範な潜在的標的に対して、自動化された脆弱性の発見と悪用により、強力なサイバー攻撃を可能にする。
③ 欺瞞や難読化の手段により、人間の制御や監視の回避を可能にする。

利用者が関連する非安全な機能を利用することを防止しようとするセーフガードとともにモデルが利用者に提供される場合であっても、この定義は適用される。

とし定義しており、これは EU の「システミック・リスクを持つ汎用目的 AI」に相当する。

その他の定義としては、本大統領令中の重要インフラは米国愛国者法（USA PATRIOT Act of 2001）の定義により、詳細は割愛するが原子炉、核物質・廃棄物（Nuclear Reactors, Materials, and Waste）も含まれる。これに対応するセクター別リスク管理機関（Sector Risk Management Agency: SRMA）は国土安全保障省（Department of Homeland Security: DHS） である。

また、最高財務責任者法（Chief Financial Officers (CFO) Act of 1990）によって米国の 24 の省庁・機関には CFO が設置されている。以下では CFO の設置されている機関を「CFO 法機関」とするが、その中には DOE、NRC も含まれる [6-44]。

(2) 原則

この大統領令は広範なものであるが、Stimers ら [6-45] によると 8 つの原則

① AI は安全、安心でなければならない。
② 責任あるイノベーション、競争、協力を促進することで、米国は AI をリードし、社会の最も困難な課題を解決する技術の可能性を解き放つことができる。
③ AI の責任ある開発と利用には、米国の労働者を支援するコミットメントが必要であ

る。

④ AI 政策は、バイデン政権の専念する平等と公民権の推進に合致したものでなければならない。

⑤ 日常生活で AI や AI を搭載した製品を使用、交流、購入する機会が増えているアメリカ人の利益は保護されなければならない。

⑥ AI が進歩し続ける中で、アメリカ人のプライバシーと市民的自由は保護されなければならない。

⑦ 連邦政府による AI の利用から生じるリスクを管理し、政府内部の能力を高めて AI の責任ある利用を規制、統治、支援し、米国人により良い結果をもたらすことが重要である。

⑧ 連邦政府は、これまでの破壊的なイノベーションと変革の時代における米国のように、グローバルな社会、経済、技術の進歩をリードすべきである。

に基づいている。本報告ではこのうち①、⑦、⑧の一部について CRS レポート [6-46] を参考に簡単に紹介する。

① AI 技術の安心・安全

NIST などに安全、安心、信頼できる AI の安全性とセキュリティを開発、展開するためのガイドラインとベストプラクティス（業界標準のコンセンサスを促進することを目的とする）、及びテスト環境を開発することを求めている。これには生成 AI 向けの AI リスクマネジメントフレームワーク [6-33] の付属リソース、及び生成 AI、デュアルユース基盤モデル向けの Secure Software Development Framework (SSDF) [6-47] の付属リソースの開発、AI 監査のためのガイダンスとベンチマークを作成するイニシアティブの立ち上げを含む（期限: 2024 年 7 月 26 日）。

a) デュアルユースAIモデルを開発する、または開発しようとしている企業は、モデルの訓練、評価とデータの所有権について政府に報告すること、b) 大規模な計算インフラを取得、開発、または所有する可能性のある企業は、計算資源の所在地と能力について政府に報告することを義務付ける（期限: 2024年1月28日）。（上記の報告義務が生じるモデルの技術条件は商務長官が定めるが、初期値としては10^{26} integer/floating-point operations （FLOPs）以上の演算で学習されたモデルや、単一のデータセンターに配置され、100 Gbits/s を超えるネットワークによって接続された、10^{20} integer/floating-point operations per second （FLOPS）以上の演算能力を有するマシンが該当する。）（注30）

（注 30） これは EU の AI 法の閾値の10^{25} FLOPs よりも一桁大きく、現時点で該当するモデルは存在しないと見られる。また、現在最速のスーパーコンピューターの演算能力は10^{18} FLOPS 程度である。

各重要インフラ分野の SRMA は、重要インフラにおける AI の導入と利用に関連する潜在的なリスクを評価・査定し、脆弱性を軽減する方法を検討し、DHS に報告する（期限: 2024 年 1 月 28 日）。また、DHS は SRMA や他の規制者との協力の下、AI RMF 及びその他の適切なセキュリティガイダンスを、重要インフラの所有者及び運営者が使用する安全及びセキュリティガイドラインに組み込む（期限: 2024 年 4 月 27 日）。

⑦ 連邦政府における AI の使用

連邦政府全体における AI の利用を調整するため、OMB 長官は議長として、連邦政府の省庁間協議会を招集する。OSTP 長官が副議長を努める。最低でも CFO 法機関の長をメンバーとして含める（注 31）（期限: 2023 年 12 月 29 日）。OMB 長官などはこの協議会の助言を参考に、AI 利用に関するガイダンス 6-47 を発行する（2024 年 3 月 28 日に発行済み）。ガイダンスの主な内容は以下の通りである。

a. 各 CFO 法機関は最高 AI 責任者（Chief Artificial Intelligence Officer: CAIO）を任命する。（期限: ガイダンスの発行から 60 日間、つまり 2024 年 5 月 27 日）

b. 各 CFO 法機関は AI ガバナンス委員会、または他の適切な機関を設置する。（期限: ガイダンスの発行から 60 日間、つまり 2024 年 5 月 27 日）

c. 人々の権利や安全に影響を与える AI の政府利用について、最低限のリスク管理方法を規定する。（期限の定めなし）

d. 各 CFO 法機関は AI 戦略を策定し、人権や安全に大きなインパクトを与える AI のユースケースを追求する。（期限の定めなし）

⑧ 政府内の AI 人材の増加

OSTP と OMB などは政府における AI 人材の急増計画を立てる。これには優先分野の特定と、本大統領令を執行するための AI 人材を含むものとする（期限: 2023 年 12 月 14 日）。

6.5.7 米国立標準技術研究所（NIST）

NIST は 2020 年国家 AI イニシアティブ法でリスクマネジメントフレームワークの策定を求められるなど、米国の AI ガバナンスで重要な役割を担っている。本節では NIST の役割のうち、AI リスクマネジメントフレームワーク（AI Risk Management Framework: AI RMF）と AI セーフティー・インスティテュートについて述べる。

(1) AI リスクマネジメントフレームワーク

（注 31） 最高 AI 責任者が指名されるまでは、各省庁の長が決定する次官補レベルまたはそれに相当する適切な職員が省庁間協議会の代表を務めるものとする、とされている。

AI リスクマネジメントフレームワーク[6-34]は 2020 年国家 AI イニシアティブ法の規定により NIST が作成したフレームワークである。このフレームワークの設計思想は、「組織や個人（AI RMF 中では AI アクターと呼ぶ）が AI システムの信頼性を高めるアプローチを身につけ、長期にわたって AI システムの責任ある設計、開発、展開、利用を促進することを支援するように設計されている。」とされている。AI リスクマネジメントフレームワークは「自発的」（voluntary）で、遵守する法的義務は無く、「全分野向け」（non-sector-specific）となっている。さらに「AI コミュニティの経験とフィードバックに基づいて更新、拡張、改善」され、「適用な可能な国際標準、ガイドライン、慣行と整合させていく」としており、日本の産業技術総合研究所の「機械学習品質マネジメントガイドライン」[6-49]などと同様に ISO 標準等への整合が想定される。

AI リスクマネジメントフレームワークは 2 部構成となっており、第 1 部では組織が AI に関連するリスクをどのように枠組化するかと、想定する説明対象について記述する。続いて AI のリスクと信頼性について分析し、信用できる（trustworthy）AI システムの特徴として、

① 妥当で信頼できる（Valid and Reliable）
② 安全（Safe）
③ 安心で弾力性がある（Secure and Resilient）
④ 説明責任があり、透明性がある（Accountable and Transparent）
⑤ 説明可能で解釈可能である（Explainable and Interpretable）
⑥ プライバシーが強化されている（Privacy-Enhanced）
⑦ 有害なバイアスを管理した上で公正である（Fair with Harmful Bias Managed）

を挙げている。第 2 部ではフレームワークの「コア」を構成し、組織が AI システムのリスクに実際に対処するための 4 つの具体的な機能

① 治める（GOVERN）
② 示す（MAP）
③ 評価する（MEASURE）
④ 対処する（MANAGE）

について記述している。今後の対応として、2023 年の大統領令「安心、安全、信頼できる人工知能の開発と利用」[6-43]で、生成 AI 向けの付属リソースの作成が指示されている。

(2) AI セーフティー・インスティテュート

「安心、安全、信頼できる人工知能の開発と利用」[6-43]では、NIST に「AI 監査のための

ガイダンスとベンチマークを作成するイニシアティブの立ち上げ」が指示されている。この指示に従って設立されたのが、AI セーフティー・インスティテュート（U. S. AISI）[6-50]とAIセーフティー・インスティテュートコンソーシアム（U. S. AI Safety Institute Consortium: AISIC）[6-51]であると見られる。AISI は英日の AI セーフティー・インスティテュートに対応する。また設立のニュースリリースの中では AISI の任務について、「ガイドラインを策定し、モデルを評価し、基礎研究を追求する」とあり、リスク評価にとどまらず、ガイドラインの策定までも行うことを示唆している。

6.6 日本の AI ガバナンス

本節では日本全体の AI ガバナンスについて概説する。日本の AI ガバナンスの構造としては分野横断的（horizontal）な原則が最上位に存在し、それを実装するためのガイダンスや標準、及び AI の安全性研究を行うセーフティー・インスティテュートなどからなる。現状ではこれらの原則、ガイダンス等に法的拘束力は無く、日本で開発・運用される AI は既存の法律、規則の枠内で規制される。ただし 2024 年 6 月 4 日に閣議決定された統合イノベーション戦略 2024[6-52]では今後、制度の在り方について検討するとしている。

6.6.1 AI 事業者ガイドライン（2024）

日本の AI ガバナンスではこれまで、2019 年 3 月に日本政府（統合イノベーション戦略推進会議[6-53]）が公表した「人間中心の AI 社会原則」[6-54]が、原則的な役割を担ってきた。それを実践するための「具体的な解説書」として、総務省による「国際的な議論のための AI 開発ガイドライン案（2017 年）」[6-55]、「AI 利活用ガイドライン（2019 年）」[6-56]、経済産業省の「AI 原則実践のためのガバナンス・ガイドライン Ver. 1.1（2022 年）」[6-57]という分野横断的で法的拘束力のないガイドラインが策定・公表されてきた。日本政府が総務省、経産省主導でこれらの 3 つのガイドラインの統合・見直しを行い、さらにこの数年の技術の発展や国内外における議論を反映して新たに策定したのが、2024 年 4 月に公表された「AI 事業者ガイドライン（第 1.0 版）」[6-58]である。

「AI 事業者ガイドライン」では、「人間中心の AI 社会原則」から

(1) 人間の尊厳が尊重される社会（Dignity）
(2) 多様な背景を持つ人々が多様な幸せを追求できる社会（Diversity & Inclusion）
(3) 持続性ある社会（Sustainability）

という 3 項目の「基本理念」を受け継いでいる。また「共通指針」として、「人間中心の AI 社会原則」で掲げられた 7 つの原則を

(1) 人間中心

(2) 安全性
(3) 公平性
(4) プライバシー保護
(5) セキュリティ確保
(6) 透明性
(7) アカウンタビリティ
(8) 教育・リテラシー
(9) 公正競争確保
(10) イノベーション

の 10 原則に整理しなおしている。また、ガイドラインの対象者を

(1) AI 開発者
(2) AI 提供者
(3) AI 利用者

の 3 者に大別し、共通の指針とそれぞれに関する事項を掲載している。さらに、共通の指針では「全ての AI 関係者向けの広島プロセス国際指針」[6-7]、「高度な AI システムを開発する組織向けの広島プロセス国際指針」[6-8] を参照するなど広島 AI プロセス包括的政策枠組みを踏まえたものとなっている。

以上のように「AI 事業者ガイドライン」では AI の開発・提供・利用について望ましい方策が示されているが、具体的にどのように実行するかまでは詳述していない。そのため別添 [6-59] において多くの実践例が示されている。

6.6.2 個別分野のガイドライン・標準

本節では日本の個別分野における AI 使用のガイドライン例について述べる。

医療分野では診断・治療で以前より AI が使われてきたが、厚生労働省は 2018 年の厚生労働省医政局医事課長の通知 [6-60] により、医師法第 17 条の「医師でなければ、医業をなしてはならない。」という法の規定と AI の関係について

- 人工知能（AI）を用いた診断・治療支援を行うプログラムを利用して診療を行う場合についても、
  – 診断、治療等を行う主体は医師である。
  – 医師はその最終的な判断の責任を負う。
  – 当該診療は医師法第 17 条の医業として行われる。

と明確化した。また、医療機器には満たすべき規格・基準が複数存在するが、それらを満

たしつつ AI を利用した医用画像診断支援システムの開発を行うためのガイドライン[6-61]が発行されている。

石油・化学プラント分野では産業技術総合研究所の「機械学習品質マネジメントガイドライン」[6-49]の第 1 版をベースに、プラント保安分野への AI の適用方法を示した「プラント保安分野 AI 信頼性評価ガイドライン」[6-62]が経産省、総務省などから発表されている。このガイドラインには「なお、本ガイドラインは法令の規定を緩和したりその解釈を示したりするものではなく、法定検査に機械学習要素を活用する場合は法定義務を遵守する必要がある。」という記述がある。また「プラント保安分野における機械学習のユースケース」として 1 章が割り当てられており、その中で

- 保全（メンテナンス）に用いる機械学習システム
  (1) 配管の肉厚予測
  (2) 配管の画像診断
  (3) 設備劣化診断
- 運転（オペレーション）に用いる機械学習システム
  (4) 異常予兆検知・診断
  (5) 運転最適化

の 5 件のユースケースが紹介されている。

自動車等の自動運転・走行分野では「自動運転・運行」に対する規制という形で法律、ガイドラインが作られている。「AI」、「人工知能」という言葉が使われていない場合も多いが、AI の使用を前提としているのは明らかである。道路交通法は 2019 年、2022 年の改正で自動運転・運行への対応が行われている（概要等[6-63, 6-64, 6-65]）。また、国土交通省が「自動運転車の安全技術ガイドライン」[6-66]を発行しているほか、農林水産省も「農業機械の自動走行に関する安全性確保ガイドライン」[6-67]を発行している。

6.6.3 AI セーフティー・インスティテュート

日本では 2024 年 2 月 14 日に AI セーフティー・インスティテュート（AISI）が情報処理推進機構（IPA）に設立された[6-68]。これは 2023 年 11 月に英国で開催された AI 安全サミットでスナク首相が設置を発表した英国の同名の組織や、米国で 2024 年 1 月に設立された同名の組織に対応するものである。経済産業省[6-69]は、暫定的な業務内容として

- 安全性評価に係る調査、基準等の検討
- 安全性評価の実施手法に関する検討
- 他国の関係機関（英米の AI セーフティー・インスティテュート等）との国際連携に関する業務

を挙げており、日本の AI の安全性評価のハブとして機能すると期待される。

6.6.4　規制のサンドボックス制度

欧米で運用されている「サンドボックス」に相当する日本の制度として、「規制のサンドボックス制度（新技術等実証制度）」[6-70, 6-71] がある。これはもともと生産性向上特別措置法（2018 年）に基づいて創設された制度で、改正産業力強化法（2021 年）により経済産業省に移管・恒久化された。この制度では「まずやってみる」ことを許容するために、期間・参加者等を限定し、既存の規制の適用を受けることなく、新しい技術・ビジネスモデルの迅速な実証を可能とするとされている。

内閣官房の資料によれば制度適用の流れは

① 新規技術による事業を検討している事業者が内閣官房の一元窓口に相談。
② 実証計画の内容を工夫し、既存の規制の適用を受けることなく実証を実施できる環境をつくる。必要があれば、規制の特例措置（新事業特例制度）を求めることも可能。
③ 実証計画を主務大臣（規制所管省庁、事業所管省庁）へ申請。（内閣官房の一元窓口がサポート）
④ 主務大臣は、実証計画が、既存の規制法令に違反しない場合には認定。主務大臣の見解は新技術等効果評価委員会（内閣府）で審議。
⑤ 実証後、規制所管省庁は、検討結果に基づき、必用な規制の撤廃又は緩和のための法制上の措置その他の措置を講じる。

となる。内閣府、経済産業省の資料には AI の事例も挙げられており、日本で AI を使用した新規事業を被規制分野で立ち上げる場合にはこの制度を利用することになると考えられる。

6.6.5　民間事業者の取組み

AI 事業者ガイドラインを含め、日本の AI ガイドラインの多くは強制力を持たない努力目標であるが、企業などが自主的にガバナンスへの取組みを行っている。AI 事業者ガイドライン別添 [6-59] ではコラムという形で NEC グループ、東芝グループ、パナソニックグループ、富士通グループという、日本の大手企業の AI ガバナンスへの取組みが紹介されている。また、NTT グループでは日本政府の法規制・ガイドラインを参照し、「NTT グループ AI 憲章」などの AI ガバナンス規程類を制定したと発表した [6-72]。NTT によると、AI 利用に伴うリスクをユースケースごとに「禁止レベル」「ハイリスク」「限定的リスク」に分類するという、リスクベースのマネージメントを行うとともに、AI に関する最高責任者（Co-Chief Artificial Intelligence Office）、AI ガバナンス室を新設するという。

日本の原子力産業界では、日立製作所 [6-73] の「社会イノベーション事業に AI を活用する

ための AI 倫理原則」、東芝[6-74]の「AI ガバナンスステートメント」、三菱電機[6-75]の「AI 倫理ポリシー」という社内ガバナンスの事例が見られる。これらはいずれも AI 利用全般を対象としており、原子力適用に特化したものではないが、「透明性・説明責任重視」 「公平性重視」「法令順守」などを謳っている。

また、社内での AI ガバナンス体制の構築には社員の AI リテラシー向上が欠かせない。AI 事業者ガイドライン別添 2A に、以下の実践例が紹介されている。

> 当社は小規模企業であり、研修対象者が少ないことから、AI リテラシーの向上の研修プログラムを自前で用意せず、外部の教材を用いることとした。米国の教育技術の営利団体である Coursera[6-76]、日本ディープラーニング協会（JDLA）[6-77]等が提供しているオンライン講座及びテキスト、経済産業省の「マナビ DX」[6-78]及び「マナビ DX Quest」[6-79]等、国内外を含めて様々な教育プログラムが利用可能である。
>
> 当社は、研修対象者の到達度を図るための JDLA の検定試験シラバスにもとづいたプログラムを活用している。JDLA の G 検定[6-80]は、AI 技術の基礎から AI 倫理まで幅広く含む内容である。また、JDLA 主催の G2023#3（2023 年 7 月 7 日実施）の G 検定合格者アンケートにおいて学習時間は 15～30 時間と答えた合格者が 3 割と多数であり、研修対象者に過度な負担にならないことも確認している。
>
> また、当社は AI も含めたデジタル技術を活用するためには社員のデジタル・リテラシー向上が必須であると考えており、「IT パスポート」[6-81]の取得を全社的に推奨している。
>
> これまで実施してきて、当社が期待している効果が出ていると思っている。例えば、AI システム・サービスのインシデントについてニュースで断片的に聞いたことがあった程度の人が、AI 技術の初歩から倫理的な側面まで習得したことで、AI のリスクにとっても当事者意識を持って考えてくれるようになった。

6.6.6 今後の展開

2024 年 6 月に閣議決定された「統合イノベーション戦略 2024」[6-52]では「② AI の安全・安心の確保」として

- イノベーション推進のためにもガードレールとなる AI 利用の安全・安心を確保するためのルールが必要である。我が国は、変化に迅速かつ柔軟に対応するため、「AI 事業者ガイドライン」に基づく事業者等の自発的な取組を基本としている。今後、AI に関する様々なリスクや、規格やガイドライン等のソフトローと法律・基準等のハードローに関する国際的な動向も踏まえ、制度の在り方について検討する。

と述べている。制度については（自発的ガバナンスと制度の検討）として具体的に

- 幅広い業種に「AI 事業者ガイドライン」の周知・浸透を図る。
- 2024 年 5 月の AI 戦略会議[6-82]で了承された「AI 制度に関する考え方」[6-83]等を踏まえ、今夏に AI 戦略会議の下で新たに開催する AI 制度研究会において、制度の在り方の検討に着手する。
- 医療、自動運転、金融等の社会への影響が大きい重要分野は、技術の進展や利用状況に応じて制度の見直しの必要性等を検討する。

とあり、2024 年 8 月に初会合が開かれた「AI 制度研究会」で AI のハードロー的規制も検討するようである。

また『「AI 制度に関する考え方」について』では「③影響大・高リスクの AI 提供者・利用者」として

> 業法・規制法がある重要インフラ等に関しては、AI 導入の基準等が必要な場合にはその法令（安全基準、設備基準等）で規制すべきと考えられる。例えば、AI を搭載した自動運転車、医療機器は、道路運送車両法、医薬品医療機器法の下で認可、承認例がある。AI の動向を踏まえ、どのような AI 利用に対して規制が必要かなど、業法ごとに検討が必要である。
>
> 業法・規制法がない重要インフラ等に関しては、技術の変化や利用状況に応じて機動的な対応が望まれるが、AI 提供者が遵守すべき事項（例えば AI 搭載製品の安全基準）については、具体的な議論の積上げが必要である。このため、法令が整備されるまでの間、及び、法令が整備された後も、運用レベルで AI 事業者ガイドラインの活用もあり得ると考えられる。

とあり、個別分野で法令等による規制検討が求められる（EU の AI 法のような包括的規制にならない）見通しである。

6.7 カナダ、韓国の AI ガバナンス

CEIMIA のホワイトペーパー[6-17, 6-84]によると、カナダと韓国では EU に近い水平的な AI 規制を法制化する動きがある。特に韓国で提案されている AI 統合法案については

> AI 統合法法案は、影響の及ぶさまざまな分野を列挙することによって、縦割りで「ハイリスク領域 AI」を定義している。「ハイリスク領域 AI」とは、「人命、身体の安全、基本的権利の保護に重大な影響を及ぼす可能性のある領域で使用される AI」と定義している。具体的には、エネルギー、飲料水、医療・機器、原子力、交通システムなどである。

と述べられており、今後、原子力分野での AI 使用が法律により規制される可能性がある。

7. まとめ

本技術ノートでは、AI の原子力分野での利用、及び規制の現状把握を目的とし、AI の基礎やリスク、リスクへの取組みも含めて概説した。

IAEA や OECD/NEA では原子力分野での AI 利用を目的として国際協力を進めている。IAEA は国際協力の成果として幅広い分野での適用を検討した報告書を発行している他、AI と革新的技術が小型モジュール炉の配備の迅速化にどのように役立つかを探るための協調研究プロジェクトの主導、AI に関する協力センターの設置など、加盟国と協力して AI 技術の開発と評価に関連する活動を主導している。OECD は一般的な AI 利用に関するガイドラインを発行しているが、原子力分野でも OECD/NEA が原子力分野での AI の利用を検討しており、加盟国の AI 利用例を調査したり、実践的なベンチマーク解析の機会を提供したりしている。

EU では包括的に AI を規制する AI 法が成立し、重大なリスクをもたらしうる原子力分野での一部の AI 利用はハイリスク AI として規制の対象となる見通しである。英国、米国は国として AI の法規制を行わず、個々の規制当局が非法的規制を行う方針である。その方針に従って、英 ONR、及び米 NRC は原子力分野での AI 規制について対応方針を示しており、数年以内に AI を規制対象とする見通しである。

原子力分野での AI 利用に関しては、原子力施設の監視・運用、設計の最適化、リスク評価などで検討が報告されている他、放射線防護・核セキュリティ、材料・構造分野での検討が進められている。しかし、日本国内では重大なリスクを判断するような利用例は見られない。また、日本では地震・津波分野への AI 適用研究が活発に行われている。

幅広い分野で AI の利用への期待が示されているものの、AI に特有のリスクへの懸念も示されている。AI のリスクに対処するための仕組みとして AI の利用に指針を与える「AI ガバナンス」が世界全体から個々の企業レベルまで導入されつつある。

参考文献一覧

2-1 統合イノベーション戦略推進会議、「人間中心の AI 社会原則」、令和 3 年 3 月, https://www8.cao.go.jp/cstp/AIgensoku.pdf（2024 年 2 月 8 日確認）

2-2 総務省、経済産業省、「AI 事業者ガイドライン（第 1.0 版）」、令和 6 年 4 月 https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html（2024 年 8 月 7 日確認）

2-3 経済協力開発機構 (OECD), “Recommendation of the Council on Artificial Intelligence”, 2019 年 5 月, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (2024 年 2 月 8 日確認)

2-4 経済協力開発機構 (OECD)、「人工知能に関する理事会勧告(総務省による非公式翻訳)」、2019 年 5 月 https://www.soumu.go.jp/mAIn_content/000642217.pdf （2024 年 2 月 8 日確認）

2-5 Turing, A. M., “Intelligent Machinery (1948)”, The Essential Turing, Oxford, pp 395-432, 2004. doi: 10.1093/oso/9780198250791.003.0016

2-6 Turing, A. M., “On computable numbers, with application to the Entscheidungsproblem”, Proceedings of the London Mathematical Society, Vol. 58, pp.230–265, 1936.

2-7 Turing, A. M., “Computing Machinery and Intelligence”, Mind, Vol. 59, pp.433–460, 1950.

2-8 Shevlin, H., Vold, K., Crosby, M, Halina, M., “The limits of machine intelligence”, EMBO reports, 2019. doi: 10.15252/embr.201949177

2-9 Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., Zhang, Y., “Sparks of Artificial General Intelligence: Early experiments with GPT-4”, arXiv, 2023. doi: 10.48550/arXiv.2303.12712

2-10 Shannon, C. E., “Programming a Computer for Playing Chess”, Philosophical Magazine, Vol. 41, pp.2-13, 1950. doi: 10.1007/978-1-4757-1968-0

2-11 Turing, A. M., “Digital computers applied to games”, Faster than Thought, 1953.

2-12 松原仁、「Deep Blue の勝利が人工知能にもたらすもの」、人工知能、12 巻、5 号、pp.698–703、平成 9 年 doi: 10.11517/jjsai.12.5_698

2-13 上野晴樹、「エキスパート・システム概論」、情報処理、28 巻、2 号、pp.147–157, 平成 9 年

2-14 Lederberg, J., “How DENDRAL was conceived and born”, A History of Medical Informatics, pp.14–44 , 1990. doi: 10.1145/89482.89484

2-15 Shortliffe, E., “Computer-based medical consultations: MYCIN” , 1976. isbn: 9780444001795

2-16 Weiss, S. M., Kulikowski, C. A., Amarel, S., Safir, A., “A modelbased method for computer-aided medical decision-making” Artificial Intelligence, Vol. 11, No.1, pp.145–

172, 1978. doi: 10.1016/0004-3702(78)90015-2

2-17 People, H. E., “The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning”, IJCAI’77, pp.1030-1037, 1977. doi: 10.5555/1622943.1623043

2-18 Erman, L. D., Hayes-Roth, F., Lesser, V. R., Reddy, D. R., “The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty”, in Readings in Artificial Intelligence, pp.349–389, 1981. doi: 10.1016/B978-0-934613-03-3.50029-5

2-19 電子計算機基礎技術開発推進委員会学術的・技術的評価ワーキング・グループ、「第五世代コンピュータ・プロジェクト最終評価報告書」、平成 5 年 http://www.jipdec.or.jp/archives/publications/J0005062（2024 年 5 月 10 日確認）

2-20 村上則夫、「わが国におけるエキスパート・システムの現状」、調査と研究、20 巻、1 号、pp.35-47、平成元年

2-21 Yokobayashi, M., Yoshida, K., Kohsaka, A., Yamamoto, M., “Development of Reactor Accident Diagnostic System DISKET Using Knowledge Engineering Technique”, Journal of Nuclear Science and Technology, Vol.23, No.4, pp.300–314, 1986. doi: 10.1080/18811248.1986.9734987

2-22 成川昇、山本孝志、佐々木則夫、「原子力プラント自動配管エキスパートシステム」、計測と制御、27 巻、10 号、pp.41–42、昭和 63 年

2-23 亀山研一、本江明、伊藤説朗、浅野明朗、「原子力プラント機器配置設計支援エキスパートシステム」、計測と制御、27 巻、10 号、pp.47–48、昭和 63 年

2-24 西山琢也、篠原靖志、「原子力発電所における類似故障・トラブルの再発防止支援用エキスパートシステムの開発」、電気学会論文誌 B、 110 巻、6 号、pp.485–494、平成 2 年 doi: 10.1541/ieeejpes1990.110.6_485

2-25 松尾豊、「人工知能は人間を超えるか」、KADOKAWA、平成 27 年 isbn: 978-4-04-080020-2

2-26 寺野隆雄、「エキスパートシステムはどうなったか?」、計測と制御、42 巻、6 号、pp.458–462、平成 15 年 doi: 10.11499/sicejl1962.42.458

2-27 Samuel, A. L., “Some Studies in Machine Learning Using the Game of Checkers”, IBM Journal, Vol.3, No.3, pp.210–229, 1959.

2-28 Raschka, S.、Mirjalili, V.、株式会社クイープ（訳）、福島真太朗（監訳）、「Python 機械学習プログラミング第 3 版」、インプレス、令和 2 年 isbn: 9784295010074

2-29 Chollet, F.、株式会社クイープ（訳）、巣籠悠輔（監訳）、「Python によるディープラーニング」、マイナビ出版、令和 4 年 isbn: 978-4-8399-7773-3

2-30 ImageNet, https://image-net.org/（2024 年 4 月 10 日確認）

2-31 ImageNet, “Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)“, https://image-net.org/challenges/LSVRC/2012/results.html, 2012 年 10 月（2024 年 4 月 10 日確認）

2-32 Fukushima, K., “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”, Biological Cybernetics, Vol.36, pp.193–202, 1980. doi: 10.1007/BF00344251

2-33 福島邦彦、「ネオコグニトロン：Deep Convolutional Neural Network」、知能と情報（日本知能情報ファジィ学会誌）、27 巻、pp.115–125、平成 27 年

2-34 Krizhevsky, A., Sutskever, I., Hinton, G. E., “ImageNet Classification with Deep Convolutional Neural Networks”, Proc. Advances in Neural Information Processing Systems 25, pp. 1090-1098, 2012.

2-35 斎藤康毅、「ゼロから作る Deep Learning 2」、オライリー・ジャパン、平成 30 年 isbn: 978-4-87311-836-9

2-36 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, I., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., “Attention Is All You Need”. arXiv, 2017. doi: 10.48550/arXiv.1706.03762

2-37 Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv, 2018. doi: 10.48550/arXiv.1810.04805

2-38 Radford, A., Narasimhan, K., “Improving Language Understanding by Generative Pre-Training”, arXiv, 2018. doi: 10.48550/arXiv.1810.04805

2-39 Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus. W., “Emergent Abilities of Large Language Models”, arXiv, 2022. doi: 10.48550/arXiv.2206.07682

2-40 斎藤康毅、「ゼロから作る Deep Learning 5」、オライリー・ジャパン、令和 6 年 isbn: 978-4-8144-0059-1

2-41 シン・アンドリュー、小川航平、「ChatGPT 大規模言語モデルの進化と応用」、リックテレコム、令和 6 年 isbn: 978-4-86594-400-6

2-42 斎藤康毅、「ゼロから作る Deep Learning 4」、オライリー・ジャパン、令和 4 年 isbn: 978-4-87311-975-5

2-43 Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, J., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis. D., “Mastering the Game of Go with Deep Neural Networks and Tree Search”, Nature, 2016. doi: 10.1038/nature16961

2-44 Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., “Mastering the Game of Go without Human Knowledge”, Nature, 2017. doi: 10.1038/nature24270

3-1 国連総会（United Nations General Assembly）, “Transforming our world: the 2030 Agenda for Sustainable Development”, 2015 年 9 月, https://sdgs.un.org/2030agenda（2024 年 6 月 21 日確認）

3-2 国連総会（United Nations General Assembly）、「我々の世界を変革する：持続可能な開発のための 2030 アジェンダ（日本語訳）」、平成 27 年 https://www.env.go.jp/earth/sdgs/（2024 年 6 月 21 日確認）

3-3 国連システム事務局長調整委員会（United Nations System Chief Executives Board for Coordination）, “A United Nations system-wide strategic approach and road map for supporting capacity development on artificial intelligence”, 2019 年 6 月, https://unsceb.org/united-nations-system-wide-strategic-approach-and-road-map-supporting-capacity-development（2024 年 6 月 21 日確認）

3-4 国際連合教育科学文化機関（United Nations Educational, Scientific and Cultural Organization: UNESCO) , “Recommendation on the Ethics of Artificial Intelligence”, 2021 年 1 月, https://unesdoc.unesco.org/ark:/48223/pf0000381137（2024 年 6 月 21 日確認）

3-5 国際電気通信連合（International Telecommunication Union: ITU), “AI for Good”, https://aiforgood.itu.int/（2024 年 6 月 21 日確認）

3-6 国際電気通信連合（International Telecommunication Union: ITU), “AI for Good Global Summit Snapshot report”, 2023, https://s41721.pcdn.co/wp-content/uploads/2021/06/SNAPSHOT-REPORT-2023-FINAL.pdf（2024 年 6 月 21 日確認）

3-7 国際原子力機関（IAEA）, “Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology”, 2023. isbn: 978-92-0-131522-9

3-8 国際原子力機関（IAEA）, “New CRP: Technologies Enhancing the Competitiveness and Early Deployment of Small Modular Reactors (I31039)”, 2022 年 7 月, https://www.iaea.org/newscenter/news/new-crp-technologies-enhancing-the-competitiveness-and-early-deployment-of-small-modular-reactors-i31039（2024 年 7 月 3 日確認）

3-9 国際原子力機関（IAEA）, “IAEA Designates First Collaborating Centre on Artificial Intelligence for Nuclear Power”, 2024 年 2 月, https://www.iaea.org/newscenter/news/iaea-designates-first-collaborating-centre-on-artificial-intelligence-for-nuclear-power（2024 年 6 月 21 日確認）

3-10 経済協力開発機構（OECD）, “Technology Foresight Forum 2016 on Artificial Intelligence (AI)”, 2016, https://www.oecd.org/st Ⅱ economy/technology-foresight-forum-2016.htm（2024 年 6 月 21 日確認）

3-11 経済協力開発機構（OECD）, “AI Intelligent machines, smart policies: Conference summary”, 2018. doi: 10.1787/f1a650d9-en

3-12 経済協力開発機構（OECD）, “Recommendation of the Council on Artificial Intelligence”, 2019 年 5 月, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 （2024 年 2 月 8 日確認）

3-13 経済協力開発機構（OECD）、「人工知能に関する理事会勧告(総務省による非公式翻訳)」、令和元年 5 月 https://www.soumu.go.jp/main_content/000642217.pdf（2024 年 2 月 8 日確認）

3-14 経済協力開発機構/原子力機関（OECD/NEA), Working Group on New Technology (WGNT), https://www.oecd-nea.org/jcms/pl_88343/working-group-on-new-technology-wgnt（2024 年 6 月 21 日確認）

3-15 経済協力開発機構/原子力機関（OECD/NEA）, Task Force on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering, https://www.oecd-nea.org/jcms/pl_77779/task-force-on-artificial-intelligence-and-machine-learning-for-scientific-computing-in-nuclear-engineering（2024 年 6 月 21 日確認）

3-16 Le Corre, J.-M., Delpe, G., Wu, X., Zhao, X., “Benchmark on Artificial Intelligence and Machine Learning for Scientific Computing in Nuclear Engineering. Phase 1: Critical Heat Flux Exercise Specifications”, https://www.oecd-nea.org/jcms/pl_89619/benchmark-on-artificial-intelligence-and-machine-learning-for-scientific-computing-in-nuclear-engineering-phase-1-critical-heat-flux-exercise-specifications（2024 年 6 月 21 日確認）

3-17 欧州議会（European Parliament）, “Artificial Intelligence Act”, 2024 年 4 月, https://artificialintelligenceact.eu/the-act/（2024 年 6 月 14 日確認）

3-18 Sovrano, F., Masetti, G., “Foreseeing the Impact of the Proposed AI Act on the Sustainability and Safety of Critical Infrastructures”, In ICEGOV ’22: Proceeding of the 15th International Conference on Theory and Practice of Electrical Governance, pp.492–498, 2022. doi: 10.1145/3560107.3560253

3-19 欧州委員会（European Commission）, “Artificial Intelligence　Questions and Answers”, 2023 年 12 月, https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683（2024 年 2 月 15 日確認）

3-20 経済協力開発機構(OECD), “Regulatory sandboxes in artificial intelligence”, 2023. doi:

10.1787/8f80a0e6-en

3-21 Veale, M., Borgesius, F. Z., “Demystifying the Draft EU Artificial Intelligence Act”. Computer Law Review International, Vol.22, No.4, pp.97–112, 2021. doi: 10.9785/cri-2021-220402

3-22 欧州連合（European Union）, “Treaty on the Functioning of the European Union”, 2009 年 1 月, http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012E/TXT&from=EN（2024 年 3 月 11 日確認）

3-23 Adelard, “The impact of AI/ML on Nuclear Regulation”, 2021 年 6 月, https://www.onr.org.uk/documents/2021/onr-rrr-121.pdf（2024 年 3 月 11 日確認）

3-24 Innovation 科学・イノベーション・技術大臣（英国）(Secretary of State for Science and Technology）, “A pro-innovation approach to AI regulation”, 2023 年 3 月. isbn: 978-1-5286-4009-1

3-25 Office for Nuclear Regulation, “ONR’s pro-innovation approach to AI regulation”, 2024 年 4 月, https://www.onr.org.uk/news/all-news/2024/04/onr-shares-pro-innovation-approach-to-regulating-aⅡn-the-nuclear-sector/（2024 年 6 月 24 日確認）

3-26 Office for Nuclear Regulation, “Artificial Intelligence”, https://www.onr.org.uk/our-expertise/innovation/artificial-intelligence/（2024 年 6 月 24 日確認）

3-27 Office for Nuclear Regulation, “Expert panel - Regulation of artificial intelligence in nuclear”, 2022 年 3 月, https://www.onr.org.uk/external-panels/artificial-intelligence.htm（2024 年 3 月 15 日確認）

3-28 Office for Nuclear Regulation, “ONR leads expert discussion on AI in nuclear industry”, 2022 年 8 月, https://news.onr.org.uk/2022/08/onr-leads-expert-discussion-on-aⅡn-nuclear-industry/（2024 年 3 月 15 日確認）

3-29 Office for Nuclear Regulation, “Expert panel sessions explore use of AI in nuclear sector”, 2023 年 6 月, https://news.onr.org.uk/2023/06/expert-panel-sessions-explore-use-of-aIIn-nuclear-sector/ （2024 年 3 月 15 日確認）

3-30 Bloomfield, R. E., Ehrenberger, W. D., “Validation and licensing of intelligent software (IAEA-CN-49/68)”, Proceedings of an International Conference on Man-Machine Interface in The Nuclear Industry (Control and Instrumentation, Robotics and Artificial Intelligence), pp.107–114, 1988. isbn: 92-0-020588-7

3-31 Robotics and Artiricial Intelligence for Nuclear (RAIN) Hub, https://rainhub.org.uk/（2024 年 3 月 12 日確認）

3-32 国立研究開発法人日本原子力研究開発機構（JAEA）、「原子力百科事典

ATOMICA」、https://atomica.jaea.go.jp/（2024 年 3 月 13 日確認）
3-33 Office for Nuclear Regulation, “Outcomes of nuclear AI regulatory sandbox pilot published”, 2023 年 11 月, https://news.onr.org.uk/2023/11/outcomes-of-nuclear-ai-regulatory-sandbox-pilot-published/（2024 年 3 月 15 日確認）
3-34 Office for Nuclear Regulation, Environment Agency, “Regulators’ Pioneer Fund (Department for Science, Innovation and Technology): Pilot of a regulatory sandbox on artificial intelligence in the nucelar sector”, 2023 年 11 月, https://www.onr.org.uk/documents/2023/onr-ea-regulators-pioneer-fund.docx（2024 年 3 月 11 日確認）
3-35 英国政府, “Open Government Licence”, https://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/uk-government-licensing-framework/open-government-licence/（2024 年 3 月 19 日確認）
3-36 Regulatory Information Conference 2023, https://www.nrc.gov/public-involve/conference-symposia/ric/past/2023/agenda.html（2024 年 3 月 14 日確認）
3-37 White, A., “Regulatory decision making on Nuclear System containing artificial intelligence”, 2023 年 1 月, https://www.nrc.gov/docs/ML2303/ML23034A191.pdf（2024 年 3 月 11 日確認）
3-38 Innovate UK, https://www.ukri.org/councils/innovate-uk/（2024 年 3 月 15 日確認）
3-39 U. S. Nuclear Regulatory Commission, “Artificial Intelligence”, https://www.nrc.gov/about-nrc/plans-performance/artificial-intelligence.html（2024 年 6 月 24 日確認）
3-40 U. S. Nuclear Regulatory Commission, “Data Science and Artificial Intelligence Regulatory Applications Workshops”, https://www.nrc.gov/public-involve/conference-symposia/data-science-ai-reg-workshops.html（2024 年 6 月 24 日確認）
3-41 Office of Nuclear Regulatory Research, “Artificial Intelligence Strategic Plan: Fiscal Years 2023-2027 (NUREG-2261)”, 2023 年 5 月, https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr2261/index.html（2024 年 6 月 17 日確認）
3-42 U. S. Nuclear Regulatory Commission, “Project Plan for the U.S. Nuclear Regulatory Commission Artificial Intelligence Strategic Plan Fiscal Years 2023-2027, Revision 0 (ML23236A279)”, 2023 年 10 月, https://www.nrc.gov/docs/ML2323/ML23236A279.pdf（2024 年 6 月 17 日確認）
3-43 U. S. Nuclear Regulatory Commission, U. S. Department of Energy, “Cooperation in the

Area of Operating Experience and Applications of Data Analytics (ML21069A328)”, 2021 年 6 月, https://www.nrc.gov/docs/ML2106/ML21069A328.pdf（2024 年 6 月 24 日確認）

3-44 U. S. Nuclear Regulatory Commission, Electric Power Research Institute Inc., “Cooperative Nuclear Safety Research (ML21263A196)”, 2021 年 6 月, https://www.nrc.gov/docs/ML2126/ML21263A196.pdf（2024 年 6 月 24 日確認）

3-45 Office of Nuclear Regulatory Research, “Exploring Advanced Computational Tools and Techniques with Artificial Intelligence and Machine Learning in Operating Nuclear Plants (NUREG/CR-7294, INL/EXT-21-61117)”, 2022 年 2 月, https://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr7294/index.html（2024 年 6 月 25 日確認）

3-46 アメリカ合衆国行政管理予算局（Office of Management and Budget）, “Guidance for Regulation of Artificial Intelligence Applications”, 2020 年 11 月, https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf（2024 年 2 月 27 日確認）

3-47 National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, 2023 年 1 月. doi: 10.6028/NIST.AI.100.1

3-48 米連邦政府人事管理局（United States Office of Personnel Management）, 「AI 力量リスト（The AI in Government Act of 2020 – Artificial Intelligence Competencies）」, 2023 年 7 月, https://www.chcoc.gov/sites/default/files/Skills-Based%20Hiring%20Guidance%20and%20Competency%20Model%20for%20Artificial%20Intelligence%20Work.pdf#:~:text=OPM%20Memorandum%2C%20The%20AI%20in%20Government%20Act%20of,to%20fill%20positions%20to%20expand%20AI%20capabilities%20government-wide. （2024 年 8 月 27 日確認）

3-49 Hanson, C. T., “Advancing Use of Artificial Intelligence at the U.S. Nuclear Regulatory Commission (ML23303A143)”, 2023 年 10 月, https://www.nrc.gov/docs/ML2330/ML23303A143.pdf（2024 年 6 月 26 日確認）

3-50 U. S. Nuclear Regulatory Commission, “Advancing Use of Artificial Intelligence at the U.S. Nuclear Regulatory Commission (SECY-24-0035)”, 2024 年 4 月, https://www.nrc.gov/docs/ML2408/ML24086A001.html（2024 年 6 月 24 日確認）

4-1 国際原子力機関（IAEA）, “Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology”, 2023. isbn: 978-92-0-131522-9

4-2 Masato Watanabe, Osamu Segawa, Yoshio Kimura, “Automatic Check System for Worker`s Protective Equipment at Entrance of Radiation Controlled Area”, SMiRT 27, 2024.

4-3 横洲弘武、田中良仁、加藤勝秀、「海洋レーダによる津波予測における AI 技術の適用」、電力土木、406 号、pp.47-50、令和 2 年

4-4 田中正暁、森健郎、岡島智史、菊地紀宏、「ARKADIA－次世代原子力プラント設計のイノベーションに向けて～原子炉構造設計最適化プロセスの実装～」、2022 年日本原子力学会春の年会, 令和 4 年 https://confit.atlas.jp/guide/event-img/aesj2022s/3J_PL04/public/pdf

4-5 国立研究開発法人　日本原子力研究開発機構、「高速炉・新型炉に関する研究開発 総合評価手法（ARKADIA）の開発」、https://www.nrc.gov/reading-rm/doc-collections/cfr/part050/part050-0065.html

4-6 日立製作所、「日立、現場データの収集技術や生成 AI を活用した「現場拡張メタバース」を開発」、令和 5 年 12 月 https://www.hitachi.co.jp/New/cnews/month/2023/12/1218.html（2024 年 8 月 9 日確認）

4-7 東芝、「大規模・複雑なプラントの状態変化の中に埋もれた異常を早期に高精度に検知する異常予兆検知 AI を開発」、令和 3 年 12 月 https://www.global.toshiba/jp/technology/corporate/rdc/rd/topics/21/2112-01.html（2024 年 8 月 9 日確認）

4-8 小田和弘、近藤誠治、谷宏幸、佐子朋生、「火力発電プラントの異常兆候検知システム」、三菱電機技報、93 巻、11 号、pp.16-19、令和元年 https://www.giho.mitsubishielectric.co.jp/giho/pdf/2019/1911105.pdf（2024 年 8 月 9 日確認）

4-9 東芝エネルギーシステムズ、「原子力事業者の安全性と信頼性に資する情報管理」、東芝レビュー、75 巻、2 号、p.41、令和 2 年 3 月 https://www.global.toshiba/content/dam/toshiba/migration/corp/techReviewAssets/tech/review/2020/02/75_02pdf/3-0.pdf（2024 年 8 月 9 日確認）

4-10 原子力エネルギー協議会（ATENA）、「原子力領域での人口知能（AI）及び先進製造技術(AMT)の活用状況について」、令和 6 年 8 月　（公開準備中）

4-11 U. S. Nuclear Regulatory Commission, Title 10, Code of Federal Regulations, Part50, “Issuance, Limitations, and Conditions of Licenses and Construction Permits” 50.65, “Requirements for monitoring the effectiveness of maintenance at nuclear power plants,

4-12 Hess, S. M., Hodges, J. L., Burr, J. M., Diven, S. W., “Use of Machine Learning to Evaluate Maintenance Rule Functional Failures at Exelon”, 2021 International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA 2021), pp.278-288, 2021.

4-13 Lin, L., Athe, P., Rouxelin, P., Avramova, M., Gupta, A., Youngblood, R., Lane, J., Dinh, N., “Digital-twin-based improvements to diagnosis, prognosis, strategy assessment, and discrepancy checking in nearly autonomous management and control system”, Annals of Nuclear Energy, Vol.166, 2022. doi: 10.1016/j.anucene.2021.108715

4-14 Tokatli, O., Das, P., Nath, R., Pangione, L., Altobelli, A., Burroughes, G., Jonasson, E. T., Turner, M. F., Skilton, R., “Robot-Assisted Glovebox Teleoperation for Nuclear Industry”, Robotics, Vol.10, 2021. doi: 10.3390/robotics10030085

4-15 Huang, Q., Peng, S., Deng, J., Zeng, H., Zheng, Z., Liu, Y., Yuan, P., “A review of the application of artificial intelligence to nuclear reactors “Where we are and what`s next””, Helyon, Vol.9, 2023. doi: 10.1016/j.heliyon.2023.e13883

4-16 Deleplace, A., Atamuradov, V., Allali, A., Pelle, J., Plana, R., Alleaume, G., “Ensemble Learning-based Fault Detection in Nuclear Power Plant Screen Cleaners”, IFAC PapersOnLine Vol.53, pp.10354-10359, 2020. doi: 10.1016/j.ifacol.2020.12.2773

4-17 Che, Y., Yurko, J., Seurin, P., Shirvan, K., “Machine learning-assisted surrogate construction for full-core fuel performance analysis”, Annals of Nuclear Energy, Vol.168, 2022. doi: 10.1016/j.anucene.2021.108715

4-18 二神敏、山野秀将、栗坂健一、氏田博士、「AI 技術を活用した確率論的リスク評価手法の高度化研究（1）AI ツールの開発計画」、日本原子力学会 2023 年春の年会、 令和 5 年 https://confit.atlas.jp/guide/event-img/aesj2023s/2C10/public/pdf?type=in

4-19 斎藤隆泰、廣瀬壮一、「波動解析や逆問題及び非破壊評価における AI・データサイエンス活用の動向」、 AI・データサイエンス論文集、4 巻、3 号、pp.852-866、令和 5 年 doi: 10.11532/jsceIII.4.3_852

4-20 Siljama, O., Koskinen, T., Jessen-Juhler, O., Virkkunen, I., “Automated Flaw Detection in Multi-channel Phased Array Ultrasonic Data Using Machine Learning”, Journal of Nondestructive Evaluation, Vol.40, No.67, 2021. doi: 10.1007/s10921-021-00796-4

4-21 Virkkunen, I., Koskinen, T., Jessen-Juhler, O., Rinta-aho, J., “Augmented Ultrasonic Data for Machine Learning”, Journal of Nondestructive Evaluation, Vol.40, 2021. doi: 10.1007/s10921-020-00739-5

4-22 Nafey, A. S., “Neural network based correlation for critical heat flux in steam-water flows in pipes”, International Journal of Thermal Sciences, Vol.48, pp.2264-2270, 2009. doi: 10.1016/j.ijthermalsci.2009.04.010.

4-23 田中浩平、「機械学習モデルによる地形情報からの工学的基盤深度の推定モデル構築」、土木学会論文集 A1（構造・地震工学）、76 巻、2 号、pp.411-423、令和 2 年

4-24　高橋幸宏、能島暢呂、香川敬生、「地震動分布のモード分解と機械学習による周期・成分別の空間特性の分析」、 土木学会論文集 A1（構造・地震工学）、78 巻、4 号、pp.478-493、令和 4 年

4-25　石井透、小穴温子、「震央方位と応答継続時間を考慮した機械学習による地点固有の地震動評価モデルの検討」、日本地震工学会論文集、22 巻、2 号、令和 4 年

4-26　小穴温子、石井透、宮下裕貴、古川慧、「強震動データベースに基づく機械学習による地震動評価モデルの構築」、日本地震工学会論文集、22 巻、4 号、pp.23-38、令和 4 年

4-27　工藤祥太、下條賢梧、溜渕功史、「1 次元畳み込みニューラルネットワークを用いた地震波形検測」、験震時報、86 巻、令和 5 年

4-28　平田直、長尾大道、「第 238 回地震予知連絡会重点検討課題「人工知能による地震研究の深化」の概要」、地震予知連絡会会報、110 巻、pp.451-452、令和 5 年

4-29　郷右近英、Post, J.、Stein, E.、Martinis, S.、Twele, Al、Mück, M.、越村俊一,「TerraSAR-X 画像の機械学習による津波被災地の自動検出」、土木学会論文集 B2（海岸工学）、69 巻、2 号、I_1441_I_1445、平成 25 年

4-30　千葉周作、Adrino, B.、Bai, Y.、越村俊一、「機械学習による事後画像のみを用いた津波被災地の建物被害の抽出」、土木学会東北支部技術研究発表会、Ⅱ-53、平成 28 年

4-31　青井真、「S-net データを用いた津波即時予測手法の開発について」、気象庁第 13 回津波予測技術に関する勉強会、平成 28 年

4-32　Makinoshima, F., Oishi, Y., Yamazaki, T., Furumura,T., Imamura, F., “Early forecasting of tsunami inundation from tsunami and geodetic observation data with convolutional neural networks”, nature communication, Vol.12, 2021. doi: 10.1038/s41467-021-22348-0

4-33　平岡伸隆、吉川直孝、伊藤和也、「深層学習による斜面表層ひずみの異常検知」、AI・データサイエンス論文集、2 巻、J2 号、pp.556-567、令和 3 年 doi: 10.11532/jsceiii.2.J2_556

4-34　鳥屋部佳苗、加村晃良、風間基樹、「強震観測データのみから地盤の液状化の程度を判定する深層学習技術の妥当性の検討 －東北地方太平洋沖地震を事例として－」、AI・データサイエンス論文集、2 巻、J2 号、pp.598-608、令和 3 年 doi: doi.org/10.11532/jsceiii.2.J2_598

4-35　桑原光平、松岡昌志、「機械学習を用いた日本全国の液状化危険度の推定」、日本地震工学会論文集、21 巻、2 号、pp.70-89、令和 3 年

4-36　Haoyang, X.、劉ウェン、丸山喜久、「機械学習に基づく 2018 年北海道胆振東部地震における斜面崩壊の推定」、土木学会論文集 A1（構造・地震工学）、78 巻、4 号（地震工学論文集第 41 巻）、 I_646-I_656、令和 4 年

4-37　奥澤康一、中岡健一、板垣昭、「ディープラーニングによる岩石の種類の判定の試み」、応用地質、63 巻、6 号、pp.291-296、令和 5 年
4-38　佐藤真俊、高橋典之、櫻井真人、相澤直之、「深層学習を用いた画像計測手法による RC 部材の地震損傷評価」、コンクリート工学年次論文集、39 巻、2 号、pp.739-744、平成 29 年
4-39　内藤昌平、門馬直一、山田哲也、下村博之、望月貫一郎、本田禎人、中村洋光、藤原広行、庄司学、「熊本地震における航空写真を用いた画像解析手法による建物被害抽出」、土木学会論文集 A1（構造・地震工学)、75 巻、4 号（地震工学論文集第 38 巻)、I_218-I_237、令和元年
4-40　藤田翔乃、畑山満則、「航空写真を用いた深層学習による地震災害時の屋根損傷家屋の把握」、土木学会論文集 D3 (土木計画学)、75 巻、6 号（土木計画学研究・論文集第 37 巻)、I_127-I_136、令和 2 年
4-41　森田高市、坂下雅信、「ディープラーニングによる RC 柱の損傷度判定に関する検討」、日本建築学会技術報告集、27 巻、66 号、pp.756-760、令和 3 年
4-42　吉岡智和、國友弘隆、「せん断破壊の特徴を深層学習させた識別機による RC 柱の損傷度推定」、コンクリート工学論文集、34 巻、pp.83-94、令和 5 年
5-1　Zhang, X., Chan, F. T. S., Yan, C., Bose, I., “Towards risk-aware artificial intelligence and machine learning systems: An overview", Decision Support Systems, Vol.159, 2022. doi: 10.1016/j.dss.2022.113800
5-2　Dastin, J.、「焦点: アマゾンが AI 採用打ち切り、「女性差別」の欠陥露呈で」, Reuters, 2018 年 10 月, https://jp.reuters.com/article/idUSKCN1ML0DM/（2024 年 5 月 22 日確認）
5-3　NIST, 「MNIST データベース (Modified National Institute of Standards and Technology database)」, 現在は http://yann.lecun.com/exdb/mnist からダウンロード可能。 （2024 年 4 月 24 日確認）
5-4　Amin, K. S., Forman, H. P., Davis, M. A., “Even with ChatGPT, race matters ", Clinical Imaging, Vol.109, 2024, doi: 10.1016/j.clinimag.2024.110113
5-5　Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., “Scikit-learn: Machine Learning in Python”, Journal of Machine Learning Research, Vol.12, pp. 2825-2830, 2011.
5-6　"Underfitting vs. Overfitting", https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html（2024 年 5 月 31 日確認）
5-7　"Gaussian Process regression: basic introductory example, https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_noisy_targets.html

#sphx-glr-auto-examples-gaussian-process-plot-gpr-noisy-targets-py（2024 年 5 月 31 日確認）

5-8 Szegedy, C., Zaremb, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., "Intriguing properties of neural networks", arXiv, 2013, doi: 10.48550/arXiv.1312.6199

5-9 Krizhevsky, A., Sutskever I., Hinton, G. E., “ImageNet classification with deep convolutional neural networks", Communications of the ACM, Vol.60, pp.84-90, 2012. doi: 10.1145/3065386

5-10 Shi, Y., Fan, C., Zou, L., Sun, C., Liu, Y., "Unsupervised Adversarial Defense through Tandem Deep Image Priors”, Electronics, Vol.9, p.1957, 2020, doi: 10.3390/electronics9111957

5-11 Eykholt, K., Evtimov, I, Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D., "Robust Physical-World Attacks on Deep Learning Visual Classification", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognitio, pp.1625-1634, 2018, doi:10.1109/CVPR.2018.00175

5-12 Papernot, N., McDaniel, P., Goodfellow, I., "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples", arXiv, 2016, doi: 10.48550/arXiv.1605.07277

5-13 Zhong. Y., Deng, W., "Towards Transferable Adversarial Attack Against Deep Face Recognition", arXiv, 2020, doi: 10.48550/arXiv.2004.05790

5-14 Creative Commons, CC BY, https://creativecommons.org/licenses/by/4.0/deed.ja（2024 年 3 月 19 日確認）

5-15 Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., Liu, Y., "Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study", arXiv, 2023, doi: 10.48550/arXiv.2305.13860

5-16 Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., Liu, Y., "Prompt Injection attack against LLM-integrated Applications", arXiv, 2023, doi: 10.48550/arXiv.2306.05499

5-17 Park, P. S., Goldstein, S., O'Gara, A., Chen, M., Hendrycks, D., "AI deception: A survey of examples, risks, and potential solutions", Patterns, Vol.5, 2024, doi: 10.1016/j.patter.2024.100988

5-18 Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A.,, Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., Zijlstra, M., "Human-level play in the game of *Diplomacy* by combining language models

with strategic reasoning", Science, Vol.378, pp.1067-1074, 2022, doi: 10.1126/science.ade9097

5-19 de Vires, A., "The growing energy footprint of artificial intelligence", Joule, Vol.7, pp.2191-2194, 2023, doi: 10.1016/joule.2023.09.004

5-20 EPRI, "Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption", 2024, https://www.epri.com/research/products/000000003002028905（2024 年 6 月 13 日確認）

5-21 Reuters, "OpenAI CEO Altman says at Davos future AI depends on energy breakthrough", 2024 年 1 月, https://www.reuters.com/technology/openai-ceo-altman-says-davos-future-ai-depends-energy-breakthrough-2024-01-16/（2024 年 6 月 13 日確認）

6-1 経済協力開発機構 (OECD), "Recommendation of the Council on Artificial Intelligence", 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449（2024 年 2 月 8 日確認）

6-2 経済協力開発機構 (OECD)、「人工知能に関する理事会勧告」 (総務省による非公式翻訳)、令和元年 https://www.soumu.go.jp/main_content/000642217.pdf（2024 年 2 月 8 日確認）

6-3 国連総会（United Nations General Assembly）, "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development”, 2024 年 3 月, https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/78/L.49&Lang=E（2024 年 6 月 14 日確認）

6-4 国際連合広報センター、「国連総会、人工知能（AI）に関する画期的な決議を採択（UN News 記事・日本語訳）」、令和 6 年 https://www.unic.or.jp/news_press/features_backgrounders/50035/（2024 年 6 月 14 日確認）

6-5 G7、「広島 AI プロセスに関する G7 首脳声明（仮訳）」、令和 5 年 10 月 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document01.pdf（2024 年 2 月 8 日確認）

6-6 G7、「広島 AI プロセス G7 デジタル・技術閣僚声明（仮訳）」、令和 5 年 12 月 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document02.pdf（2024 年 2 月 8 日確認）

6-7 G7、「全ての AI 関係者向けの広島プロセス国際指針（仮訳）」、令和 5 年 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document03.pdf（2024 年 2 月 8 日確認）

6-8 G7、「高度な AI システムを開発する組織向けの広島プロセス国際指針（仮訳）」、令和 5 年 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document04.pdf（2024 年 2 月 8 日確認）

6-9 G7、「高度な AI システムを開発する組織向けの広島プロセス国際行動規範（仮訳）」、令和 5 年 https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document05.pdf（2024 年 2 月 8 日確認）

6-10 High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI ", 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai（2024 年 2 月 15 日確認）

6-11 欧州議会（European Parliament）, "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI", 2023 年 12 月, https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai（2024 年 2 月 15 日確認）

6-12 欧州議会（European Parliament）, "Artificial Intelligence Act: MEPs adopt landmark law", 2024 年 3 月, https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law（2024 年 6 月 14 日確認）

6-13 欧州連合理事会（Council of the EU）, "Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI", 2024 年 5 月, https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/pdf/（2024 年 6 月 14 日確認）

6-14 日本貿易振興機構 (JETRO)、「EU、AI を包括的に規制する法案で政治合意、生成型 AI も規制対象に」、令和 5 年 12 月 https://www.jetro.go.jp/biznews/2023/12/8a6cd52f78d376b1.html（2024 年 2 月 15 日確認）

6-15 欧州議会（European Parliament）, "Artificial Intelligence Act", 2024 年 4 月, https://artificialintelligenceact.eu/the-act/（2024 年 6 月 14 日確認）

6-16 European Parliamentary Research Service, "Artificial intelligence Act", 2023 年 6 月, https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792（2024 年 2 月 15 日確認）

6-17 Huw Roberts, Marta Ziosi, Cailean Osborne, Lama Saouma, " A Comparative Framework for AI Regulatory Policy", The International Centre of Expertise on Artificial Intelligence

in Montreal, 2023, https://www.ceimia.org/wp-content/uploads/2023/05/a-comparative-framework-for-ai-regulatory-policy.pdf（2024 年 7 月 18 日確認）

6-18 Michael Veale, Frederik Zuiderveen Borgesius, "Demystifying the Draft EU Artificial Intelligence Act", Computer Law Review International, Vol. 22, pp.97-112, 2021, doi: 10.9785/cri-2021-220402

6-19 Maria Webb, "Insights: Breaking Down the Transformative Journey of GPT Models in AI, from GPT-1 to GPT-4", Technopedia, 2023, https://www.techopedia.com/gpt-series-evolution-insights（2024 年 2 月 16 日確認）

6-20 Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W., "Emergent Abilities of Large Language Models", arXiv, 2022, doi: 10.48550/arXiv.2206.07682

6-21 European Commision（欧州委員会）, "Commission establishes AI Office to strengthen EU leadership in safe and trustworthy Artificial Intelligence", 2024 年 5 月, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2982（2024 年 8 月 28 日確認）

6-22 Roberts, H., Babuta, A., Morley, J., Thomas, C., Taddeo, M., Floridi, L., "Artificial intelligence regulation in the United Kingdom: a path to good governance and global leadership?", Internet Policy Review, Vol.12, 2023, doi: 10.14763/2023.2.1709

6-23 経済協力開発機構 (OECD), "National AI policies & strategies", https://oecd.ai/en/dashboard/overvies（2024 年 2 月 21 日確認）

6-24 Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., Perrault, R., "The AI Index 2023 Annual Report", AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2023, https://hai.stanford.edu/research/ai-index-2023（2024 年 2 月 21 日確認）

6-25 科学・イノベーション・技術大臣（英国）(Secretary of State for Science, Innovation and Technology, "A pro-innovation approach to AI regulation", 2023 年 3 月, isbn: 978-1-5286-4009-1

6-26 Centre for Data Ethics and Innovation（英国）, “The roadmap to an effective AI assurance ecosystem”, 2021 年 12 月, https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem（2024 年 8 月 21 日確認）

6-27 科学・イノベーション・技術大臣（英国）(Secretary of State for Science, Innovation and Technology, "Introducing the AI Safety Institute", 2023 年 11 月, isbn: 978-1-5286-4538-6

6-28 科学・イノベーション・技術省（英国）(Department for Science, Innovation and Technology), "Introduction to AI assurance", 2024 年 2 月, https://www.gov.uk/government/publications/introduction-to-ai-assurance（2024 年 2 月 26 日確認）

6-29 科学・イノベーション・技術省（英国）(Department for Science, Innovation and Technology), "Portfolio of AI assurance techniques", 2023 年 6 月, https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques（2024 年 2 月 26 日確認）

6-30 アラン・チューリング研究所（Alan Turing Institute）, https://www.turing.ac.uk/about-us（2024 年 2 月 26 日確認）

6-31 AI 標準ハブ（AI Standards Hub）, https://www.aistandardshub.org/the-ai-standards-hub/（2024 年 2 月 26 日確認）

6-32 Neowin, “UK begins to draft AI regulations focusing on the most powerful language models", 2024 年 4 月, https://www.neowin.net/news/uk-begins-to-draft-ai-regulations-focusing-on-the-most-powerful-language-models/（2024 年 6 月 17 日確認）

6-33 The White House, "AI 権利章典の青写真 (Blueprint for an AI Bill of Rights)", 2022 年 10 月, https://www.whitehouse.gov/ostp/ai-bill-of-rights/（2024 年 2 月 27 日確認）

6-34 NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", 2023 年 1 月, doi: 10.6028/NIST.AI.100.1

6-35 NIST AI リスクマネジメントフレームワーク（AI セーフティー・インスティテュートによる日本語翻訳）, 2024 年 7 月, https://aisi.go.jp/2024/07/04/ai_nist_rmf_ja_news/ (2024 年 8 月 9 日確認)

6-36 The White House, "Maintaining American Leadership in Artificial Intelligence (Executive Order 13859)", 2019 年 2 月, https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence（2024 年 2 月 27 日確認）

6-37 Harris, L. A., "Artificial Intelligence: Background, Selected Issues, and Policy Considerations", Congressional Research Service, 2021 年 9 月, https://crsreports.congress.gov/product/pdf/R/R46795（2024 年 2 月 27 日確認）

6-38 アメリカ合衆国行政管理予算局（OMB）, "AI アプリケーション規制ガイダンス（Guidance for Regulation of Artificial Intelligence Applications）", 2020 年 11 月, https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf（2024 年 2 月 27 日確認）

6-39 The White House, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (Executive Order 13960)", 2020 年 12 月, https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government（2024 年 2 月 27 日確認）

6-40 アメリカ合衆国議会（United States Congress）, "National Artificial Intelligence Initiative Act of 2020 (Division E)", 2021 年, https://www.congress.gov/bill/116th-congress/house-bill/6216（2024 年 2 月 28 日確認）

6-41 アメリカ合衆国議会（United States Congress）, "AI in Government Act of 2020 (Division U, Title I)", 2021 年, https://www.congress.gov/bill/116th-congress/house-bill/2575（2024 年 2 月 27 日確認）

6-42 内閣府 科学技術・イノベーション推進事務局, 「米国の AI 権利章典（AI Bill of Rights）について」, 2022 年 12 月, https://www8.cao.go.jp/cstp/ai/ningen/r4_2kai/siryo3.pdf（2024 年 2 月 27 日確認）

6-43 The White House, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Order 14110)", 2023 年 10 月, https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence（2024 年 2 月 27 日確認）

6-44 最高財務責任者会議（Chief Financial Officers Council）, https://www.cfo.gov/members（2024 年 3 月 1 日確認）

6-45 Stimers, P., Serafino, M. C., Roberson, J. E., Wooten, T., Barsky, D. J., "What to Know About the New Artificial Intelligence Executive Order", Holland & Knight LLP., 2023 年 10 月, https://www.hklaw.com/en/insights/publications/2023/10/what-to-know-about-the-new-artificial-intelligence-executive-order（2024 年 2 月 29 日確認）

6-46 Harris, L. A., "Highlights of the 2023 Executive Order on Artificial Intelligence for Congress", Congressional Research Service, 2023 年 8 月, https://crsreports.congress.gov/product/pdf/R/R47843（2024 年 2 月 27 日確認）

6-47 NIST, "Secure Software Development Framework (SSDF)", 2022 年 2 月, doi: 10.6028/NIST.SP.800-218

6-48　アメリカ合衆国行政管理予算局（OMB）, "M-24-10 Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence", 2024 年 3 月, https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf（2024 年 2 月 27 日確認）
6-49　産業技術総合研究所、「機械学習品質マネジメントガイドライン 第 4 版」、令和 5 年 12 月。https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html（2024 年 2 月 9 日確認）
6-50　U. S. Artificial Intelligence Safety Institute, https://www.nist.gov/aisi（2024 年 8 月 9 日確認）
6-51　Artificial Intelligence Safety Institute Consortium (AISIC), https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic（2024 年 8 月 9 日確認）
6-52　内閣府、「統合イノベーション戦略 2024」、令和 6 年 6 月 https://www8.cao.go.jp/cstp/tougosenryaku/2024.html（2024 年 6 月 19 日確認）
6-53　内閣府、「統合イノベーション戦略推進会議」、https://www8.cao.go.jp/cstp/tougosenryaku/kaigi.html（2024 年 2 月 8 日確認）
6-54　統合イノベーション戦略推進会議、「人間中心の AI 社会原則」、令和 3 年 3 月 https://www8.cao.go.jp/cstp/aigensoku.pdf（2024 年 2 月 8 日確認）
6-55　AI ネットワーク社会推進会議、「国際的な議論のための AI 開発ガイドライン案」、平成 29 年 7 月 https://www.soumu.go.jp/main_content/000499625.pdf（2024 年 2 月 8 日確認）
6-56　AI ネットワーク社会推進会議、「AI 利活用ガイドライン」、令和元年 8 月 https://www.soumu.go.jp/main_content/000637097.pdf（2024 年 2 月 8 日確認）
6-57　AI 原則の実践の在り方に関する検討会、AI ガバナンス・ガイドライン WG、「AI 原則実践のためのガバナンス・ガイドライン Ver.1.1」、令和 4 年 1 月 https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20220128_report.html（2024 年 2 月 8 日確認）
6-58　総務省, 経済産業省、「AI 事業者ガイドライン (第 1.0 版)」、令和 6 年 4 月 https://www.meti.go.jp/press/2024/04/20240419004/20240419004-1.pdf（2024 年 6 月 19 日確認）
6-59　総務省、経済産業省、「AI 事業者ガイドライン (第 1.0 版) 別添（付属資料）」、令和 6 年 4 月 https://www.meti.go.jp/press/2024/04/20240419004/20240419004-1.pdf（2024 年 6 月 19 日確認）
6-60　厚生労働省医政局医事課長、「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」、平成 30 年 12 月

https://www.mhlw.go.jp/content/10601000/000468150.pdf（2024 年 2 月 9 日確認）

6-61 経済産業省、国立研究開発法人日本医療研究開発、「医用画像診断システム（人工知能を利用するものを含む）開発ガイドライン２０１９(手引き)」、令和元年 12 月
https://www.meti.go.jp/policy/mono_info_service/healthcare/iryou/downloadfiles/pdf/47_guideline.pdf（2024 年 2 月 9 日確認）

6-62 石油コンビナート等災害防止 3 省連絡会議（経済産業省、総務省消防庁、厚生労働省）、「プラント保安分野 AI 信頼性評価ガイドライン 第 2 版」、令和 3 年 3 月
https://www.meti.go.jp/press/2020/03/20210330002/20210330002.html（2024 年 2 月 9 日確認）

6-63 警察庁、「道路交通法の一部を改正する法律（概要）」、令和元年 12 月
https://www.npa.go.jp/bureau/traffic/selfdriving/trafficact.pdf（2024 年 6 月 20 日確認）

6-64 警察庁、「道路交通法の一部を改正する法律（概要）」、令和 4 年 4 月
https://www.npa.go.jp/bureau/traffic/selfdriving/L4-summary.pdf（2024 年 6 月 20 日確認）

6-65 警察庁交通局交通企画課自動運転企画室長、「特定自動運行に係る許可制度の創設について」、令和 4 年 6 月
https://www.mlit.go.jp/jidosha/content/001485116.pdf（2024 年 6 月 20 日確認）

6-66 国土交通省自動車局、「自動運転車の安全技術ガイドライン」、平成 30 年 9 月
https://www.mlit.go.jp/common/001253665.pdf（2024 年 6 月 20 日確認）

6-67 農林水産省、「農業機械の自動走行に関する安全性確保ガイドライン」、令和 5 年 3 月 https://www.maff.go.jp/j/press/nousan/gizyutu/attach/pdf/230329-2.pdf（2024 年 6 月 20 日確認）

6-68 独立行政法人情報処理推進機構、「プレス発表 AI セーフティー・インスティテュートを設立」、令和 6 年 2 月
https://www.ipa.go.jp/pressrelease/2023/press20240214.html（2024 年 2 月 14 日確認）

6-69 経済産業省、「AI セーフティー・インスティテュートを設立しました」、令和 6 年 4 月 https://www.meti.go.jp/press/2023/02/20240214002/20240214002.html（2024 年 2 月 14 日確認）

6-70 内閣官房 新しい資本主義実現本部事務局（規制のサンドボックス 政府一元的窓口）、「規制のサンドボックス制度（新技術等実証制度）について」
https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/underlyinglaw/sandboximage516.pdf（2024 年 2 月 14 日確認）

6-71 経済産業省 経済産業政策局 産業創造課 新規事業創造推進室、「産業競争力強化法に基づく事業者単位の規制改革制度について」 https://www.meti.go.jp/policy/jigyou_saisei/kyousouryoku_kyouka/shinjigyo-kaitakuseidosuishin/220715_sankyouhou_kiseikaikaku_gaiyou.pdf（2024 年 2 月 14 日確認）

6-72 日本電信電話株式会社、「NTT グループの AI ガバナンス規程類の制定、及び AI ガバナンスの推進体制について」、令和 6 年 6 月 https://group.ntt/jp/newsrelease/2024/06/07/240607a.html（2024 年 6 月 7 日確認）

6-73 吉田順、柳井孝介、「社会イノベーション事業のための日立の AI 倫理原則とその実践」、日立評論、令和 3 年 https://www.hitachihyoron.com/jp/archive/2020s/2021/sp/sp01/index.html（2024 年 8 月 9 日確認）

6-74 東芝、「東芝グループ AI ガバナンスステートメント」、令和 4 年 8 月 https://www.global.toshiba/jp/technology/corporate/ai-statement.html（2024 年 8 月 9 日確認）

6-75 三菱電機、「三菱電機グループ「AI 倫理ポリシー」策定」、令和 3 年 12 月 https://www.mitsubishielectric.co.jp/news/2021/pdf/1215-a.pdf（2024 年 8 月 9 日確認）

6-76 Coursera, https://www.coursera.org/（2024 年 6 月 20 日確認）

6-77 一般財団法人 日本ディープラーニング協会（JDLA）、https://www.jdla.org/（2024 年 6 月 20 日確認）

6-78 独立行政法人 情報処理推進機構（IPA）、「マナビ DX」。 https://www.jdla.org/certificate/general/（2024 年 6 月 20 日確認）

6-79 独立行政法人 情報処理推進機構（IPA）、「マナビ DX クエスト」。 https://dxq.manabi-dx.ipa.go.jp/index.html（2024 年 6 月 20 日確認）

6-80 一般財団法人 日本ディープラーニング協会（JDLA）、「G 検定」。 https://www.jdla.org/certificate/general/（2024 年 6 月 20 日確認）

6-81 独立行政法人 情報処理推進機構（IPA）、「IT パスポート試験」。 https://www3.jitec.ipa.go.jp/JitesCbt/index.html（2024 年 6 月 20 日確認）

6-82 内閣府、「AI 戦略会議 第 9 回」、令和 6 年 5 月 https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/9kai.html (2024 年 6 月 20 日確認)

6-83 AI 戦略チーム、「「AI 制度に関する考え方」について」、令和 6 年 5 月 https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo2-1.pdf（2024 年 6 月 20 日確認）

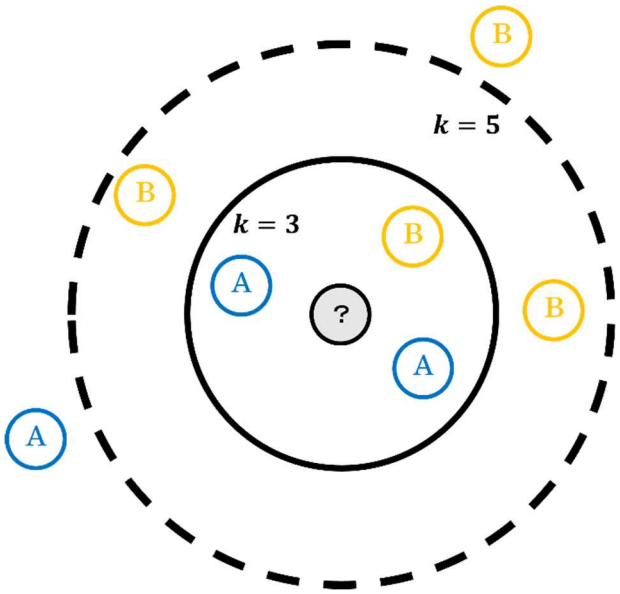
6-84 Centr' d'Expertise International de Montréal en Intelligence Artificielle (CEIMIA), "A

Comparative Framework for AI Regulatory Policy: Phase 2", 2024, doi: 10.5281/zenodo.12575144

執筆者一覧

原子力規制庁　長官官房　技術基盤グループ　システム安全研究部門
　宮﨑　利行　主任技術研究調査官
　野口　法秀　技術研究調査官
原子力規制庁　長官官房　技術基盤グループ　地震・津波研究部門
　太田　良巳　主任技術研究調査官
　東　喜三郎　主任技術研究調査官
原子力規制庁　長官官房　技術基盤グループ　シビアアクシデント研究部門
　梁田　勇太　技術研究調査官

付録１　機械学習モデルの例



図付 1　k 近傍法の例

Fig. app 1 An example of k-nearest neighbor method.

注）この例では「?」のノードは k=3 の時「A」に、k=5 の時「B」に分類される。

In this example, the “?” node is classified as “A” when k=3 and “B” when k=5.

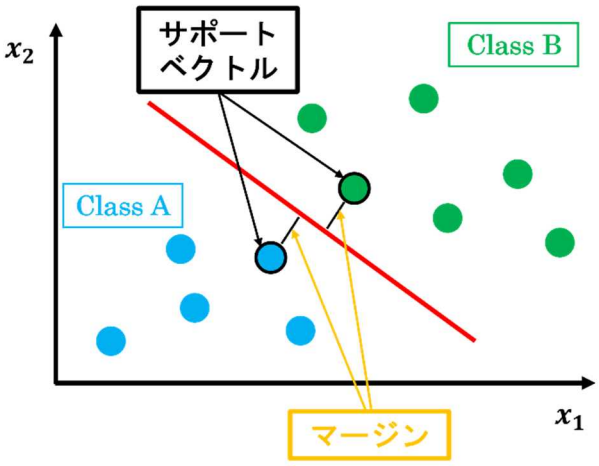
(1)　k 近傍法（k-nearest neighbor: k-NN）

既に分類したデータ（学習データ）を覚えておいて、新たに分類する場合には最も近い k 個のデータと同じ分類を行うという k 近傍法（k-NN）の例を図付 1 に示す。この例では k=3 の時、「?」のノードに最も近い 3 つのノードでは「A」が 2 個、「B」が 1 個で「A」の方が多いため「?」は「A」と分類される。同様に k=5 の時は近傍に「B」が多いため、「B」に分類される。k の値が小さいとノイズの影響を受けやすくなること、偶数だと同数で分類できなくなる場合があることから、一般には k=3、5 程度の値が用いられることが多い。k 近傍法は学習時には計算を行う必要は無いことから「怠惰学習」にも分類される。しかし全ての学習データを記憶しておく必要があり、分類時に距離の計算を行う必要があることから、特に学習データの規模が大きい場合にメモリー消費が大きくなり、予測速度も遅くなる。比較的古くから存在する手法だが、手間をかけずにある程度の性能が得られるので機械学習を手始めに適用する手法としては現在も有効である。

(2)　サポートベクターマシーン（Support Vector Machine: SVM）

学習データから、各データ点とのマージンが最大となる超平面を求めることで分類を行う手法である（回帰へも適用できる）。サポートベクターマシーンの例を図付 2 に示す。こ

の例では Class A と Class B の「マージン」が最大になるような線を引いて判別を行っている。それぞれの Class で境界線に最も近い点をサポートベクター（サポートベクトル）と呼ぶ。カーネル関数を用いることにより、非線形な分類へも適用することができる。機械学習の代表的な手法であり、広く用いられていたこともあったが、現在は決定木系のモデルの人気が高いようである。



図付 2　サポートベクターマシーンの概念図

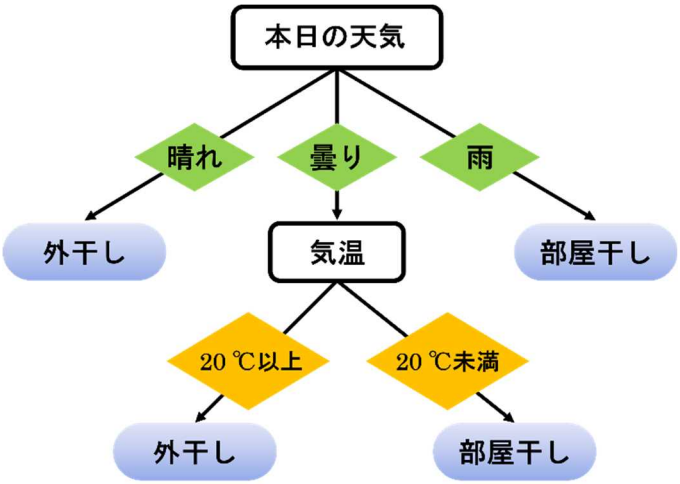
Fig. app.2 Example of a support vector machine

注）この例では Class A と Class B の「マージン」が最大になるような線（赤）を引いて判別を行っている。それぞれの Class で境界線に最も近い点をサポートベクター（サポートベクトル）と呼ぶ。

In this example, a red line is drawn to maximize the "margin" between Class A and Class B for discrimination. The point closest to the boundary line in each Class is called the support vector.

(3) 決定木

決定木は学習データセットをその変数にしたがって樹状に分類する手法である。図付 3 に決定木の例を示す。この例では「本日の天気」と「気温」によって行動を決定している。予測データも決定木にしたがって分類していくが、その分類方法は容易に解釈可能であるため、解釈性の高い分類手法であるとみなされている。一方、分類に使用する変数を増やしていくと、その項目に当てはまるデータ数が少なくなるため学習データに適合しすぎるという過剰適合（過学習）に陥りやすいという問題点がある。



図付 3　決定木の例

Fig. app.3 An example of a decision tree. In this example

注）この例では「本日の天気」と「気温」によって行動を決定している。

The action is determined by "today's weather" and "temperature".

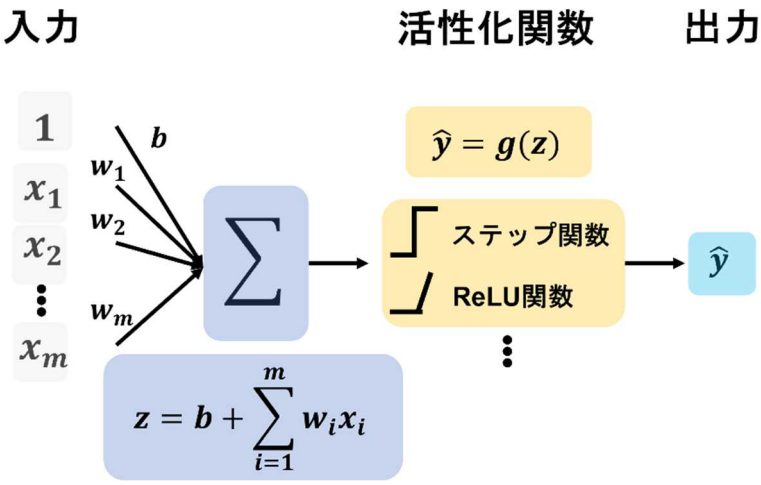
純粋な決定木は過学習に陥りやすいという問題があるが、その問題点を解消するために提案された手法がランダムフォレスト[付 1]である。ランダムフォレストではランダムに選択した学習データから、ランダムに選択した特徴量で作成した決定木を複数用意し、その多数決で予測を行う。予測の解釈は難しくなるが、過学習が起こりにくいため比較的高い性能が得られ、必要な計算量もそれほど大きくないため機械学習の代表的手法として用いられることが多い。

ランダムフォレストでは比較的浅い決定木を複数用意して多数決で分類を行ったが、浅い決定木による分類結果から、誤差に応じて重みをつけて決定木を学習する、という過程を繰り返して複数の決定木を作り、その多数決で予測を行うのが勾配ブースティング決定木である。代表的なものとしては Gradient Boosting Decision Tree(GBDT) [付 2]、AdaBoost [付 3]などがある。ランダムフォレストと比較すると予測性能が高い場合が多いが、計算時間が長くなる傾向がある。近年では XGBoost [付 4]、LightGBM [付 5]など、高い予測性能を保ちながらも計算時間を短縮した予測器が登場し、シャローラーニングに分類される機械学習手法の中でも使われる機会が増えているようである。

(4) ニューラルネットワーク（多層パーセプトロン）

パーセプトロン、あるいは人工ニューロンは Rosenblatt が 1957 年に考案し、1958 年に論文発表[付 6]した、人間の脳の機能をモデル化したアルゴリズムである。一般的には、入力（スカラー、あるいはベクトル）をスカラー値にアフィン変換し、そのスカラー値を非線形の活性化関数で変換した結果が閾値以上なら 1 を、閾値より小さければ-1（あるいは 0）を出力するようなモデルである。一般に、活性化関数としてはステップ関数、シグモイド

関数、tanh 関数、ランプ関数（Rectified Linear Unit: ReLU）などが用いられる。パーセプトロンの例を図付 4 に示す。この例では入力x_1~x_mに対して重みw_1~w_mをかけ、バイアスbと共に足し合わせ、非線形な活性化関数を通して出力している。一層のパーセプトロンにより線形分離可能な問題を解くことができる。



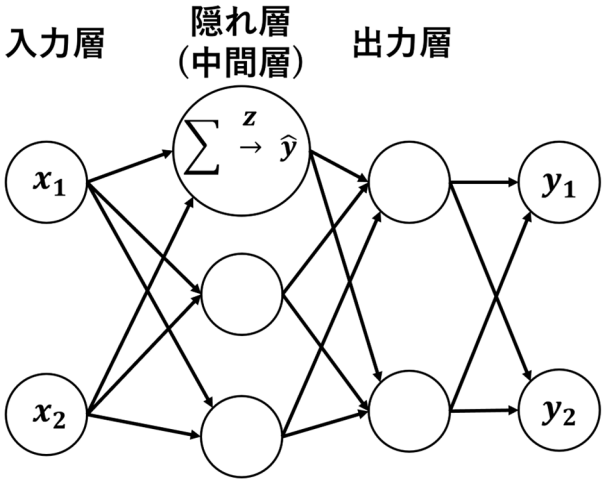
図付 4　パーセプトロンの例

Fig. app.4 An example of a perceptron

注）入力x_1~x_mに対して重みw_1~w_mをかけ、バイアスbと共に足し合わせる。ステップ関数、ReLU 関数などの非線形な活性化関数を通して出力する。

Multiply inputs x_1~x_m by weights w_1~w_m and add them together with bias b. Output through a nonlinear activation function such as a step function, ReLU function, etc.

また Hornik ら[付 7]は、入力層、隠れ層、出力層の三層から構成される多層パーセプトロン（ニューラルネットワーク）は任意の連続関数を任意の精度で近似できるという、いわゆる普遍近似定理（universal approximation theorem、万能近似定理などともいう）を示した。この定理より、ニューラルネットワークを任意の分類、回帰問題に適用することができる。図付 5 に三層ニューラルネットワークの例を示す。この例では入力層、隠れ層（中間層）、出力層の三層になっており、入力x_1、x_2に対しy_1、y_2を出力する。三層以上で構成されるニューラルネットワークの学習法はバックプロパゲーション（誤差逆伝搬法）と呼ばれる方法が 1967 年に甘利[付 8]によって考案され、1986 年の Rumelhart ら[付 9]によって再発見された。バックプロパゲーションによって三層以上のニューラルネットワークの学習が容易になり、広く使われるようになった。



図付 5　三層ニューラルネットワークの例

Fig. app.5 An example of a three-layer neural network

注）この例では入力層、隠れ層（中間層）、出力層の三層になっており、入力x_1、x_2に対しy_1、y_2を出力する。

In this example, there are three layers: an input layer, a hidden layer (intermediate layer), and an output layer, which outputs y_1, y_2 for the inputs x_1, x_2.

単純な関数近似だけであれば三層のニューラルネットワークで充分であるが、層を増やせばそれだけ多くの情報をニューラルネットワークに記憶させることができる。現在主流で使われている深層学習（ディープラーニング）は数十~数百層のニューラルネットワークが用いられており、三層、あるいは数層のニューラルネットワークとは質的に異なると考えられている。そこで、特に層数の大きいニューラルネットワークによる学習を深層学習（ディープラーニング）と呼んで区別しているが、具体的に何層以上が深層学習かという共通認識が存在しているわけではない。深層学習も含むニューラルネットワークの詳細については、例えば斎藤[付 10]を参照されたい。

付録 2　説明可能な AI

機械学習モデル、特に深層学習モデルは動作がブラックボックスに近く、なぜモデルがそのような予測をしたのか、その予測は信頼できるのか、といった疑問に答えることが難しい。そこで AI の説明（解釈）可能性を高めるための研究が行われてきており、それらの総称として説明可能な AI（eXplAInable AI: XAI）という言葉が使われている(一例として Gunning ら[付 11])。これは XAI という特定の AI モデルが開発されているわけでなく、AI の説明可能性を高めるための様々なアプローチを総称として XAI としている。なお、AI の中でもルールベースで推論を行うエキスパートシステムや、機械学習の中でも線形回帰を用いた手法、純粋な決定木法などは比較的説明性が高い。

機械学習の XAI として比較的よく使われているものに、LIME、SHAP と呼ばれる手法がある。Ribeiro ら[付 12] が提案した、局所的に解釈可能なモデル不可知論的説明（Local Interpretable Model-Agnostic Explanations: LIME）は、複雑な機械学習モデルを、局所的に解釈可能なモデルで近似することにより解釈する。モデル不可知論的[注 32]なので、元のモデルがどのようなものであっても適用可能である。

Lundberg ら[付 13, 付 14]の提案した、SHapley Additive exPlanations (SHAP)と呼ばれる手法は、Sharpley[付 15]の提案したシャープレー値（Sharpley value）と呼ばれる協力ゲーム理論の方法を利用している。シャープレー値は複数のプレイヤーが協力してゲームをプレイして報酬を得た場合にその報酬を公平に分配する方法で、機械学習の各特徴量をプレイヤーに見立てることで、予測への貢献度を評価する。SHAP では個々の予測に対する各特徴量の貢献度が求められるとともに、モデルのマクロ的な予測性能に対する各特徴量の貢献も求めることができ、機械学習の XAI 手法としては広く用いられている。

また画像認識の分野では、Zhou ら[付 16]の提案した CNN がどこに着目して判断したのかを可視化する Class Activation Map (CAM)と呼ばれる手法や、Selvaraju ら[付 17]が改良した Grad-CAM などの手法が提案されている。

一般に、このようなモデルの予測に対する説明可能性（explainability）、あるいは予測の意味を理解するための解釈可能性（interpretability）と、モデルの予測性能は相反関係にあると考えられており、実際そのような図を掲載している例もある（例として Gunning ら[付 11]の Figure 1）。それが、「AI モデルの性能を高めるにはモデルが複雑になり、その複雑なモデルを説明可能にするための仕組みが必要である」という考え方につながっているが、それに対する反論も存在する。Rudin ら[付 18, 付 19]は、あるデータセットに対して「羅生門セット」という十分精度の高いモデルが多数存在し、その中には解釈可能なモデルが少なくとも一つは存在する、という主張をしている。「羅生門セット」は「羅生門効果」（黒澤明の映画「羅生門」に由来する）という、一つのデータセットに対して誤差が最小程度になる

[注 32]　モデルについての知識は無いということ。

複数の説明（モデル）が存在するという効果に基づいている。

例えば 0 から 9 までの手書きの数字を画像化し、正解の数字がラベルとして与えられている MNIST というデータセット（訓練用データが 60000 枚、評価用データが 10000 枚）は 1998 年にリリースされて以来標準的なデータセットの一つとして AI モデルの評価に使われてきたが、Baldominos ら[付 20]によると、現在では誤認識率が 1%を切るモデルが多数存在し、それらが「羅生門セット」となっている。MNIST の数字認識に関しては大きな「羅生門セット」が存在している状態なので、これ以上複雑なモデルを導入することに現実的な意味は無い。

付録 3　コンピューターシミュレーションへの適用

物理モデルなどに従って現実の問題を数値計算で再現・模擬するコンピューターシミュレーションは、一般に多くの計算リソース（CPU、計算時間）を要求する。より少ない計算コストで、シミュレーションと同様の計算結果を得たいという需要は常に存在し、そのために開発・使用されているのが代理モデル（surrogate model、サロゲートモデル）である。代理モデルはメタモデル（metamodel）などとも呼ばれ、Razavi ら[付 21]によれば、大まかにデータ駆動的にシミュレーションモデルの応答を模擬する応答局面代理モデル（response surface surrogate model）と、物理モデルを簡略化した低忠実度代理モデル（lower-fidelity）の 2 系統が存在するが、ここでは前者について述べる。データ駆動的にシミュレーションモデルの応答を模擬するというのは機械学習、特に普遍近似定理により任意の連続関数を近似できるニューラルネットワークと相性が良く、多項式や基底関数を用いる方法と並んで 2000 年代から検討が進められていた。性能の向上やライブラリの整備に伴って、ニューラルネットワークや深層学習を代理モデルに使用する例が増えているようである（例えば Agarwal ら[付 22]、Hürkamp ら[付 23]）。

一方、機械学習を使った代理モデルには、必ずしも入力に対して物理的に正しい出力が出るわけではないこと、前提条件のない状態からモデルを訓練するのは無駄が多いという問題がある。そこで Raissi ら[付 24]は、ニューラルネットワークの学習時に支配方程式の制約を加えるという Physics-informed neural networks (PINNs)と呼ばれる方法を提案した。PINNs に関しては現在も研究が進められており、日本でも、例えば増田ら[付 25]が浅水伝播シミュレーションでの検討例を報告している。一方で実用化には課題が多く、今後検討を進めなければならない部分が多い。

流体の運動を計算的に求める数値流体力学（computational fluid dynamics; CFD）は、原子力に限らず幅広い産業分野で使用される、利用範囲の広いシミュレーションである。一方、高精度の流体シミュレーションは非常に高い計算資源を要求し、特に、典型的な用途の一つで予測精度を高めることに大きな経済価値のある気象シミュレーションではその時代の最先端のスーパーコンピューターが使われてきた。従って AI を流体シミュレーションに適用して、精度を維持しながら必要な計算資源を減らす、あるいは同等の計算資源でより高い精度を実現することは社会的・経済的に大きなインパクトがあり、PINNs などの適用例が報告されている（例えば Sun ら[付 26]や Cai ら[付 27]）。また、非圧縮性流体に限るが、Wandel ら[付 28, 付 29]の提案している、教師なし学習でナビエストークス方程式を解くことができるニューラルネットワークは、膨大な時間をかけて代理モデルの訓練用データを用意することなく、流体シミュレーションの速度を大幅に短縮できるとしている。

数値流体力学の中でも、大気の状態変化を予測する数値予報（天気予報）は社会的・経済的影響が大きく、膨大な計算資源を必要とするため精力的に研究が進められている（例

えば Schulz ら[付 30])。一部では AI による予想が従来の数値予報の精度を上回ったという報告もなされている（例えば Espeholt ら[付 31]、Bi ら[付 32])。また数値流体力学への適用ではないが、生成 AI を用いて数値予報の計算負荷を減らすという報告もなされている。Lorenz [付 33]が指摘したように、数値予報、特に長期予報では初期値のわずかなずれが、最終的には大きな予測結果の違いをもたらす。したがって気象庁 [付 34] を含む各国の気象予報機関は、長時間・長期予報では初期値をわずかに変えた同条件のシミュレーションを複数（気象庁の場合は 5~51 メンバー）実行する「アンサンブル予報」を行っている（この場合の「アンサンブル」は AI で使われる「アンサンブル」手法とは全く異なるので注意が必要である)。アンサンブル予報では規模の大きい（多くの場合は地球全体）シミュレーションを数十通り実行するので、非常に大きな計算資源を必要とする。そこで Li ら[付 35] は過去の予報を学習し、生成 AI で 2 通りのシミュレーションから 31 通りの予報を生成するシステムを開発したとしている。さらに、2024 年 6 月にマイクロソフト[付 36] は地球大気の大規模基盤モデル「Aurora」を発表した。Bondnar ら[付 37] によると、Aurora は従来の数値予報と比較すると約 5000 倍高速（数値予報は 352×36 コアの CPU を使用しているのに対し、Aurora は GPU が A100 一枚の構成）でありながら、多くのケースで従来の数値予報を上回る精度を得ているという。また大気中の汚染物質の拡散シミュレーション例も示しており、大気の基盤モデルにより放射性汚染物質の拡散を高速で高精度に予測できる可能性がある。

参考文献（付録）

付1 Leo, B., “Random Forests”, Machine Learning, Vol.45, pp.5–32, 2001. doi: 10.1023/A:1010933404324

付2 Friedman, J. H., “Stochastic gradient boosting”, Computational Statistics & Data Analysis, Vol.38, No.4, pp.367–378, 2002. doi: 10.1016/S0167-9473(01)00065-2

付3 Freund, Y., Schapire, R. E., “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”, Journal of Computer and System Sciences, Vol.55, No.1, pp.119-139, 1997. doi: 10.1006/jcss.1997.1504

付4 Chen, T., Guestrin, C., “XGBoost: A Scalable Tree Boosting System”, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794, 2016. doi: 10.1145/2939672.2939785

付5 Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., “LightGBM: A Highly Efficient Gradient BoostingDecision Tree”, Advances in Neural Information Processing Systems, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf（2024 年 5 月 10 日確認）

付6 Rosenblatt, F., “The perceptron: A probabilistic model for information storage and organization in the brain”, Psychological Review, Vol.65, No.6, pp.386-408, 1958.

付7 Hornik, K., Stinchcombe, M., White, H., “Multilayer feedforward networks are universal approximators”, Neural Networks, Vol. 2, No.5, pp.359-366, 1989. doi 10.1016/0893-6080(89)90020-8

付8 Amari, S., “Dreaming of mathematical neuroscience for half a century”, Neural Networks, Vol. 37, pp.48-51, 2013. doi: 10.1016/j.neunet.2012.09.014

付9 Rumelhart, D. E., Hinton, G. E., Williams, R. J., “Learning representations by back-propagating errors”, Nature Vol. 323, pp.533–536, 1986. doi: 10.1038/323533a0

付10 斎藤康毅、「ゼロから作る Deep Learning」、オライリー・ジャパン、平成 28 年 isbn: 978-4-87311-758-4

付11 Gunning, D., Vorm, E., Wang, J. Y., Turek, M., ”DARPA's explainable AI (XAI) program: A retrospective”, Applied AI Letters, Vol.2, 2021. doi: 10.1002/ail2.61

付12 Ribeiro, M. T., Singh, S., Guestrin, C., "Why Should I Trust You?": Explaining the Predictions of Any Classifier”, 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016. doi: 10.1145/2939672.2939778

付13 Lundberg, S. M., Lee, S.-I., “A Unified Approach to Interpreting ModelPredictions”, 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 1090-1098, 2017.

付14　Lundberg, S. M., Erion, G. G., Lee, S.-I., ”Consistent Individualized Feature Attribution for Tree Ensemble”, arXiv, 2018. doi: 10.48550/arXiv.1802.03888

付15　Shapley, L. S., ”A value for n-person games”, Contribution to the Theory of Games, Vol.0, pp.307-317, 1953.

付16　Zhou, B, Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., ”Learning Deep Features for Discriminative Localization”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921-2929, 2016. doi: 10.1109/CVPR.2016.319

付17　Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., ”Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”, 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017. doi: 10.1109/ICCV.2017.74

付18　Rudin, C., ”Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”, Nature Machine Intelligence, Vol.1, 2019. doi: 10.1038/s42256-019-0048-x

付19　Semenova, D., Rudin, C., Parr, R., ”On the Existence of Simpler Machine Learning Models”, arXiv, 2019. doi: 10.48550/arXiv.1908.01755

付20　Baldominos, A., Saez, Y., Isasi, P., ”A Survey of Handwritten Character Recognition with MNIST and EMNIST”, Applied Sciences, Vol.9, 2019. doi: 10.3390/app9153169

付21　Razavi, S., Tolson, B. A., Burn, D. H., ”Review of surrogate modeling in water resources”, Water Resources Research, Vol. 48, 2012. doi: 10.1029/2011WR011527

付22　Agarwal, S., Tosi, N., Breuer, D., Padovan, S., Kessel, P., Montavon, G., ”A machine-learning-based surrogate model of Mars’ thermal evolution”, Geophysical Journal International, Vol. 222, pp.1656-1670, 2020. doi: 10.1093/gji/ggaa234

付23　Hürkamp, A., Gellrich, S., Dér, A., Herrmann, C., Dröder, K., Thiede, S., ”Machine learning and simulation-based surrogate modeling for improved process chain operation”, The International Journal of Advanced Manufacturing Technology, Vol. 117, pp. 2297-2307, 2021. doi: 10.1007/s00170-021-07084-5

付24　Raissi, M., Perdikaris, P., Karniadakis, G. E., ”Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”, Journal of Computational Physics, Vol. 378, pp. 686-707, 2019. doi: 10.1016/j.cp.2018.10.045

付25　増田和輝、金澤剛、「Physics-Informed Neural Networks による浅水波伝播シミュレーションに関する基礎研究」、AI・データサイエンス論文集、4 巻、pp. 26-35, 令和 5 年 doi: 10.11532/jsceiiii.4.3_26

付26　Sun, L., Gao, H., Pan, S., Wang, J.-X., “Surrogate Modeling for Fluid Flows Based on Physics-Constrained Deep Learning Without Simulation Data”, Computer

Methods in Applied Mechanics and Engineering, Vol. 361, 2020. doi: 10.1016/j.cma.2019.112732

付27 Cai, S., Mao, Z., Wang, Z., Yin, M., Karniadakis, G. E., “Physics-informed neural networks (PINNs) for fluid mechanics: a review”, Acta Mechanica Sinica, Vol. 37, pp. 1727-1738, 2022. doi: 10.1007/s10409-021-01148-1

付28 Wandel, N., Weinmann, M., Klein, R., “Unsupervised Deep Learning of Incompressible Fluid Dynamics”, arXiv, 2020. doi: 10.48550/arXiv.2006.08762

付29 Wandel, N., Weinmann, M., Klein, R., ”Teaching the incompressible Navier-Stokes equations to fast neural surrogate models in three dimensions”, Physics of Fluids, Vol. 33, 2021. doi: 10.1063/5.0047428

付30 Schulz, M. G., Batancourt, C., Gong, B., Kleinert, F., Lagguth, M., Leufen, L. H., Mozaffari, A., Stadtler, S., ”Can deep learning beat numerical weather prediction?”, Philosophical Transactions of the Royal Society A, Vol. 379, 2021. doi: 10.1098/rsta.2020.0097

付31 Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., Bell, A., Kalchbrenner, N., ”Deep learning for twelve hour precipitation forecasts”, nature communications, Vol. 13, 2022. doi: 10.1038/s41467-022-32483-x

付32 Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian,Q., “Accurate medium-range global weather forecasting with 3D neural networks”, Nature, Vol. 619, 2023. doi: 10.1038/s41586-023-06185-3

付33 Lorenz, E. N., “Deterministic Nonperiodic Flow”, Journal of the Atmospheric Sciences, Vol. 20, pp.130-141, 1963. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2

付34 気象庁情報基盤部、「令和 5 年度数値予報解資料集」、令和 6 年 issn: 2758-1330

付35 Li, L., Cavert, R., Lopez-Gomez, I., Sha, F., Anderson, J., “Generative emulation of weather forecast ensembles with diffusion models”, Science Advances, Vol. 10, pp. 533-538, 2024. doi: 10.1126/sciadv.adk4489

付36 Microsoft, “Introducing Aurora: The first large-scale foundation model of the atmosphere”, 2024 年 6 月, https://www.microsoft.com/en-us/research/blog/introducing-aurora-the-first-large-scale-foundation-model-of-the-atmosphere/ (2024 年 6 月 7 日確認)

付37 Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A., Gupta, J. K., Tambiratnam, K., Archibald, A., Heider, E., Welling, M., Turner, R. E., Perdikaris, P., “Aurora: A Foundation Model of the Atmosphere”, arXiv, 2024. doi: 10.48550.arXiv.2405.13063