ISSN 2436-9462

SBI金融経済研究所 所報
vol. 6
2024.8

SBI Research Review

次世代・デジタル金融の社会デザインを考える

SBI 金融経済研究所は、先端テクノロジーを活用した次世代・デジタル金融およびその市場のあり方を検討し、戦略的な提言を発信してまいります。
提言を通じて、日本社会全体のより良い発展に貢献することを目指します。


SBI
Financial and Economic Research Institute

SBI Research Review

SBI金融経済研究所 所報 vol.6 2024.8

CONTENTS

巻頭言

政井 貴子｜SBI 金融経済研究所 理事長

政井 貴子

SBI 金融経済研究所 理事長
1965 年生まれ。トロント・ドミニオン銀行、クレディ・アグリコル銀行、新生銀行などにて金融市場関連業務を推進し、2013 年新生銀行初の女性執行役員に就任、2016 年日本銀行審議委員に任命される。2021 年より現職。

2021 年に設立された SBI 金融経済研究所は、多くの皆様のお力添えのお陰を持ちまして、無事開所 4 年目を迎えることができました。

最初に始めた研究所の活動は、日本を代表する方々のデジタル金融に関する知見の共有でした。2021 年末の岩村充上席研究顧問（早稲田大学名誉教授）のご寄稿を端緒として、アカデミア、実務家そして法律家の皆さまにさまざまな視点からデジタルを切り口とした多くのご寄稿・論考をお寄せいただき、設立まもない研究所の活動を支えていただきました。この場をお借りして、改めて全ての皆様に御礼申し上げます。

さて、この 3 年間をデジタル金融という切り口で振り返りますと、その話題の移ろいや実装の速さを実感します。例えば、2021 年の 3 月にコロナ禍の最中に行われた FIN/SUM の冒頭のパネルディスカッションのテーマは、「暗号資産ビジネスの可能性と未来」でした。その後、暗号資産市場は、暗号資産そのものとは無縁の理由で幾つかの大きなスキャンダルに見舞われ、米国ではこれらに関連してデジタルバンクランという預金者行動や銀行の清算を経験することになりました。杉浦研究主幹（当時）は、寄稿において、2001 年のエンロン事件と共通する金融・資本市場における発行体の信頼担保の重要性を指摘しています。また、天谷知子氏（元金融庁金融国際審議官、現農林中金総合研究所エグゼクティブアドバイザー）は、伝統的な金融市場と新たなデジタル資産との相関の高まりを指摘し、適切なリスク評価整備の必要性を述べています。本年に入り、米国で暗号資産を原資産とした ETF が上場され、新たな展開が始まりました。本号では、斉藤賢爾早稲田大学大学院教授が論考を寄せてくださっています。当研究所は、デジタル資産に対する意識の変化を追跡するため、2022 年より「次世代金融に関する一般消費者の関心と利用度に関するアンケート調査」を開始しました。これまでの調査では、消費者が新しいデジタル資産に対して慎重な態度を示していることが観測されています。一方で、本年実施する第 3 回目のアンケートでは、新たな動きに伴い、消費者心理に変化の兆しが見られるのか注目しています。

そして、2024 年の FIN/SUM における最初のパネルディスカッションテーマは、「Generative AI が変える金融機関の未来」でした。また、開催期間を通じて、AI に関連したテーマが多く取り上げられていました。つい 1 年前の 2023 年には、AI が人類を凌駕するというような漠然とした話題が多かったように記憶していますが、約 1 年後の今、私自身が AI と対話しながら文章を構成したり、簡易な分析や比較を行ったりすることが可能になりました。この

ように新たな技術を具備したサービスが急速に消費者に受け入れられていることを実感します。代表的なのは、金融サービスが一般サービスに埋め込まれている「Embedded Finance」ですが、中でも BaaS（Banking as a Service）の広がりは、金融に対する信頼と、消費行為における信頼、言い換えればブランド力という我々利用者が元来無意識に区別していた意識が何らかの理由で薄まった結果と受け止めています。こうした新たな展開の背景には、大きなマクロ経済環境の変化があることも見逃せません。所報第５号の巻末対談では、超低金利環境という変化が金融界に危機感を持たせ、将来を見据えた動きを加速させた役割があったと指摘しています。

一方、課題も浮き彫りになってまいりました。今年の FIN/SUM でも、「ホールセール決済の将来像」、「送金の未来」等決済に関わるテーマが取り上げられました。PayPal などの決済手段の広がりにより、P2P の場面では国境を越えた決済の利便性が向上しましたが、日本においては、証券決済やクロスボーダー決済などまだ多くの課題が残っています。デジタル化の進展を機会として捉え、日本の金融市場が世界で存在感を増すためには、既存の枠組みにとらわれない我が国の金融市場の一層の深化と進化が欠かせません。

こうした問題意識をもとに、当研究所では昨年末に山沖特任研究員を座長とする「次世代金融インフラの構築を考える研究会」を立ち上げ、７月に今後の指針を公表いたしました。今後は、公表した指針に基づき、目指すべき姿を提言してまいります。

また、デジタル金融の変遷にはマクロ金融経済環境の変化も大きく関わっていることから、金融のデジタル化を内包するさらに大きなテーマである長期的な経済社会システムについて、竹中平蔵名誉理事長を座長に「2040 年の経済社会研究会」を新たに発足させ、発信を始めました。

これらの研究会を軸に知識を一層深めつつ、日本経済への貢献に資する提言を行うために、既存の枠組みにとらわれず新たな視点を提示し、問題意識を喚起してまいりたいと思います。引き続きご支援を賜りますようお願い申し上げ、巻頭のご挨拶とさせていただきます。

解題

副島 豊｜SBI 金融経済研究所 研究主幹

本号では金融のインフラを取り上げた。一口に金融インフラといっても、その範囲は広い。法規制や金融制度、IT インフラ、金融業や市場・決済システムの産業構造、監督・モニタリング、金融業の企業ガバナンスや金融システムのガバナンス、各種協会や自主規制団体、関連産業（会計・法律事務所やコンサルティング、信用情報機関、通信・データ産業）、情報やルールの標準化、市場慣行やインテグリティ、金融業界の文化など、様々なものが金融インフラを作り上げている。そして、その先進性や機能度が金融サービス全体の質や、ひいては経済成長、国民経済厚生にも影響してくる。

そうした多様な金融インフラの構成要素のなかから、本号では 3 つのテーマについて 4 つの論文を収録した。デジタルアイデンティティに関する柴田・﨑村論文と中山論文、分散型金融システムと伝統的金融システムの結節点に関する斉藤論文、生成 AI に関する副島論文である。また、SBI 金融経済研究所は昨年 12 月より「次世代金融インフラの構築を考える研究会」を設置しており、金融インフラの課題や将来像に関する様々な議論を行っている。本号には 7 月 5 日に公表した最初の報告書を掲載し、研究会メンバーによるとりまとめ後の座談会の模様を紹介している。

- 柴田・﨑村論文「デジタルアイデンティティを巡る世界の潮流」
- 中山論文「「欧州デジタル ID 枠組み規則」制定の経緯と欧州デジタル ID ウォレットの共通仕様」
- 斉藤論文「伝統的金融に呑まれる分散型金融」
- 副島論文「生成 AI ウォークスルー：基本技術、LLM、アプリケーション実装」
- 次世代金融インフラの構築を考える研究会報告書、「次世代金融インフラの構築を考えるに当たっての指針」

【柴田・﨑村論文、中山論文】

デジタルアイデンティティは、デジタル化社会における身元確認（Identity Proofing & Verification）と当人認証（Authentication）を支えるインフラ基盤であり、個人をオンラインで識別するために使用される一連の属性や情報の集合体である。これにはユーザー名、パスワード、生体認証データのほか、位置情報やネットの閲覧など様々な個人行動の記録も含まれうる（個人以外にも組織やデバイスも識別対象に含まれうる）。デジタルアイデンティティは、特定のエンティティが誰であるかを確認し、その信頼性を保証する基盤を提供するものである。それゆえ、金融サービスに止まらずデジタル社会のあらゆるサービスに共通した重要インフラとなっている。

自分固有の情報や設定を必要とするようなサービスを利用する場合、「自分が誰であるのか」をサービス提供主体に示して登録してもらう必要がある。その真実性や本人情報の量や精度は、サービス内容によって大きく異なる。SNSであれば実名でなく仮名で登録可能なものが多い。各種サービスは、業法によって、あるいは自主規制によって、どのような本人情報を必要とするか、その身元確認保証レベル（Identity Assurance Level）は大きく異なっている。モバイルフォン購入の際に求められる身元確認保証レベルは時代とともに変遷を遂げてきており、現在の電子マネーについても、身元確認が不要なものもあれば、そうでないものもある。

一般に、金融サービスの申込においては、運転免許証やマイナンバーカードのような公的な身元確認が求められ、そのうえで当人認証手段が提供される。デジタルサービスの一例として、銀行預金口座からの送金をスマートフォンアプリで行う場合を考えてみよう。銀行はインターネットの向こう側から送られてきた送金指示が、確かに口座保有者として登録済みである貴方からの指示であることを確認する必要がある。これが当人認証であり、その前提として身元確認を伴った本人登録が必要となる。

デジタルアイデンティティの発行主体は公的部門に止まらず、民間企業もサービス利用者に対して様々なデジタルアイデンティティを発行し、サービス提供時の認証や認可に活用している。エコシステムプレーヤーのみならず個人向け小規模サービス業においても、マーケティング上、顧客管理は重要である。こうしたデジタルアイデンティティの活用シーンにおいて、技術発展によって可能となってきたことがある。個人の属性や情報の集合体であるデジタルアイデンティティを、サービス利用に際して選択的に提供する機能である。例えば、年齢確認だけが求められるサービス提供においては、氏名や住所や年齢すらも示す必要がない。その年齢制限をクリアしているという事実だけを証明することができれば十分である。コロナワクチンの接種、資格や学位の取得など、0か1で事実が判明すれば事足りるものは少なくない。

優れたサービスを享受する（提供する）ためにはデジタルアイデンティティの適切な活用が重要である一方、プライバシーの確保においてもデジタルアイデンティティの管理が必須となる。本号掲載の最初の２本の論文、柴田・﨑村論文と中山論文は、こうしたデジタルアイデンティティの最前線の動向を論じたものである。

柴田・﨑村論文は、主要国のオープンバンキングの現状を概説したうえで、プライバシー保護、金融犯罪防止、相互運用性確保といったデジタルアイデンティティの重要な役割を論じている。後半では、Verifiable Credential（VC）やデジタルアイデンティティウォレットといった具体的な技術動向を紹介し、EU を中心とした国際的な動きと課題を展望している。VC とは、個人の属性情報に対する証明書であり、信頼できる機関によって発行されたデジタル情報である。暗号技術によって保護されており、必要な属性情報だけを開示することを可能にする技術が活用されている。また、同論文は、デジタル庁によるスマートフォンでのマイナンバーカード提示や EU との協力覚書締結といった日本の取り組みを紹介し、デジタルアイデンティティ技術の進展がもたらす未来について論じている。

中山論文は、先行する EU を取り上げ、欧州デジタル ID 枠組み規則の改定と欧州デジタル ID ウォレットの仕様を解説したものである。EU では、加盟国間での電子識別手段の相互運用性を確保し、異なる国で発行された電子 ID が相互に認識され利用可能となるよう、eID と呼ばれる電子識別手段が 2014 年に採択された（eIDAS 規則）。その後、適用が推進されてきたが、各国間の足並みが揃わず民間のサービスへの利用普及が妨げられていた。そこで、本年 5 月に eIDAS 規則を改定し、加盟国は 2 年以内に欧州デジタル ID ウォレットを市民に提供する義務を負うことになった。同ウォレットは ID や属性情報を管理するアプリケーションであり、運転免許証や医療処方箋、教育資格、電子署名機能などデジタルアイデンティティに関連する様々な情報を扱うことができる。標準化を進めるためにシステム開発のツールボックス仕様を定めたほか、エコシステムとして機能するよう発行事業者や検証サービス事業者、登録機関、監督機関、認定機関など多数の関連主体を取り込んだ役割設定が工夫されている。

デジタルアイデンティティは、優れた金融サービスを創造し提供していく際に、その機能や競争優位性の確保において重要な金融インフラの一つとなっている。これら２本の論文を通じて世界の潮流と先進的なグランドデザインを知ることは、日本の金融経済・社会厚生にかかる国家戦略を考えるうえで極めて有益である。特に、中山論文に目を通した読者は、社会全体での実装と活用をどう進めていくかにおいて、日本の国家戦略が立ち遅れていることを強く意識することになろう。

【斉藤論文】

金融インフラにおいて、近年の世界的なトレンドを生み出しているのが分散型金融システムであり、その中核にあるのが分散台帳技術である。当研究所の所報や Web レポートでも、分散台帳技術がどのような金融インフラ変革をもたらしつつあるかを多く取り上げてきた。暗号資産やセキュリティトークン、ステーブルコイン、DeFi、NFT などである。また、金融規制当局や中央銀行が、こうした動きにどのような対応を行ってきたか、あるいは自ら活用しようとしているのかも金融インフラの展開を考えるうえで重要であり、金融規制や CBDC を巡る動向として取り上げてきた。例えば、この 1 年を振り返ると以下のような研究所の論文・Web レポートがあげられる（所報掲載は * 印）。

- 「次世代デジタル金融の実現に向けパブリックチェーン技術に回帰する金融機関の最新動向」、石川大紀、2023 年 12 月
- 「セキュリティトークン最新事情と将来展望：2024 年夏」、朏仁雄、2024 年 7 月
- 「トークン化がもたらす金融システムの未来と軌跡」*、齊藤達哉、2024 年 3 月
- 「RWA の現状、今後と法規制」、斎藤創、2024 年 4 月
- 「セキュリティトークンの現状と課題」、村松健、2023 年 12 月
- 「セキュリティトークン市場の現状と将来像」*、平田公一・政井貴子、2023 年 8 月
- 「NFT の持続可能性を考える ― NFT は長期保存できる資産と言えるのか？―」、中山靖司、2024 年 3 月
- 「NFT は本当に「唯一無二」と言えるのか？― NFT の信頼性を高める一つの方法の提案―」、中山靖司、2023 年 11 月
- 「暗号資産等クロニクル」米国版、EU・UK 版、国際機関版等、2023 年
- 「金融システムの未来像を探る中央銀行の挑戦」*、副島豊、2024 年 3 月
- 「CBDC のオフライン決済を巡る議論」、小早川周司、2023 年 8 月
- 「暗号資産の機能と国際的規制アプローチの変遷」、天谷知子、2023 年 12 月
- 「米国における暗号資産規制の動向」*、中山靖司、2023 年 8 月
- 「米国における暗号資産規制を巡るもうひとつの論点 ― DeFi（分散型金融）をどうするか―」*、湯山智教、2023 年 8 月
- 「米リップル社を巡る裁判の略式判決を読み解く ―個人向けに販売される XRP は、なぜ「有価証券」ではないのか―」、中山靖司、2023 年 8 月
- 「資金決済法・金融商品取引法の改正経緯から紐解くデジタル決済手段（暗号資産類似）等の定義と注目点」*、山沖義和、2023 年 8 月
- 「ステーブルコイン法制の６つの勘所」*、河合健、2023 年 8 月
- 「令和６年度税制改正（暗号資産関連）の振り返り」、安河内誠、2024 年 1 月

また、金融インフラを考えるうえでの歴史的視点の重要性（例えば、江戸期の分散分権型金融システムから明治の中央集権型システムへの移行）も意識しており、下記のような論文レポートを公表している。

- 「幕末維新期日本の貨幣制度と貨幣使用の変遷 －デジタル通貨時代における複数通貨の併存と統合を見据えて－」*、鎮目雅人、2023 年 8 月
- 「江戸時代の金融イノベーション 1 －証券取引市場の形成－」、高槻泰郎、2024 年 7 月

これらに共通したテーマが、性質が異なる 2 つのシステム、すなわち分散分権型システムと中央集権型システムをどのように併存させ、あるいは新旧システムを接続させるかという視点である。これは、IT インフラとしての視点だけでなく、ガバナンス、法規制、監督モニタリング実行、金融システム安定、利用者保護、AML/CFT、金融産業構造といった金融インフラを構成する様々な視点において共通のテーマとなっている。例えば、KYC の実現には、暗号資産交換所という分散型システムと伝統的金融システムの結節点になる組織が重要な役割を果たしている。これは、サトシ・ナカモトが描いた世界には存在していない構図である。金融機関や中央銀行の分散台帳技術への接近も、プライベートチェーンという、これまたサトシ・ナカモトの世界観には存在しなかったものを創り出した。そして、上記にあげた石川レポートは、プライベートチェーンを司る金融機関や金融インフラ企業が、パブリックチェーンとの接続を求めてクロスチェーン技術に強い関心を持ち始めている最近の動きを紹介している。

斉藤論文は、分散型の新しい金融システムといえども、伝統的な金融システム無くしては成り立たないという基本的理解をまず整理している。そのうえで、最近の米国暗号資産 ETF や日本における合同会社 DAO の認可の動きは、両者の接近というよりは、むしろ分散型システムが伝統的金融システムに呑み込まれる動きであるという見方を示している。「分散を志しながらも、伝統的金融との接点を通してしか自らを維持できない台帳システム自体に問題の根幹はあると言えるのかもしれない」という論文の指摘は、決してネガティブな批判ではない。最後に示されている言説、すなわち、分散型台帳技術の利点をもう一度振り返りながら、伝統的金融には依存しない、別の動作条件の下で機能する台帳システムの登場を志向する時期が来ているというのが筆者の本意である。これは、IT インフラやガバナンス、法規制といった金融インフラの様々な側面において、2 つのシステムをどう併存させるか、連結させるかという金融システム全体の再設計に関わってくるプロポーザルである。同様な視点は、「次世代金融インフラの構築を考える研究会」の報告書にも登場してくる。

【副島論文】

この 1 ～ 2 年、金融産業のみならず、社会経済活動全般を大きく変えていく基盤技術として注目を集めているのが生成 AI である。少なくとも金融業界において、現在のイベントで最も人が集まるのは生成 AI を取り上げたカンファランスやワークショップであり、これが DX やビッグデータ、情報利活用、IT システム開発の新潮流などと連結して、大きなうねりを作りだしている。ChatGPT が登場した 1 年半ほど前に少し触って離れてしまったきりの読者は、現在の最新サービスに触れてほしい。長大な PDF 文書ファイルを一瞬で読み込み、質問に応じて的確自在に回答するという現在の生成 AI のフロンティアを無償で体験することができる。GPT 以外にも多くの大規模言語モデルが登場しており、その知識量や言語生成能力の劇的な向上を目の当たりにすることができる。

生成 AI の意義は 2 つあると考えられる。1 つ目は、言葉という人間の最も基本的なコミュニケーション手段をコンピュータに対して直接投げかけることが可能になった点である。これまでは、プログラムを書いて指示する、あるいは誰かが作ったプログラムやアプリケーションを用いることで操作が行われてきた。ボイスコマンドやチャットボットという前駆的なものが登場していたが、その汎用性において全く次元が異なるものとなっている。言葉で指示を出し、文章や画像、音声、プログラムほか種々の情報を生成することが可能となっている。数年前までは SF の世界でしか成立しえなかった技術である。

2 つ目は、言語生成モデルが知識や情報のデータベースとして機能するようになった点である。ニューラルネットワークをベースとする言語モデルは、巨大な文書のデータベース、例えば、インターネットからの大量の収集文や著作権が切れた膨大な書籍群から言語のパターン性を学習する。言語生成モデルの基本構造は極めてシンプルであり、こういう文脈で言葉が並んでいたら次はこの言葉が適切であろうという推論を一語ごとに繰り返すことで文章が作成されていく。このような方法で人間並みの文章生成が可能になるとは考え難いが、急速な技術発展により短期間のうちに実現してしまった。もちろん、上述のような言語生成モデルの成り立ちは、必然的にハルシネーション（幻覚、もっともらしい嘘）を伴う。しかし、言語生成モデルが大規模化し、学習データセットが巨大になってくると、文の展開や文章のパターン性に内包されている情報が知識として言語モデルの中に取り込まれていった。現在の大規模言語モデルは、様々な分野の専門家並みの知識を既に獲得しており、更に特化した情報などを追加学習させることが可能となっている。

生成 AI 技術が加速的に進化する一方で、多くの金融機関が生成 AI の活用に苦慮している。金融業は、情報の生産および情報の処理を行う産業であり、その膨大な情報は、数値データのみならず文章データとして生産され、処理されている。それゆえ、金融業における生成 AI とりわけ大規模言語モデルの

活用余地は非常に大きい。しかし、情報セキュリティの課題や IT システムの開発運用方針、専門人材の不足などの課題に直面し、顧客向けの金融サービスや内部業務への応用は徐々にしか進展していない。また、ビジネスの現場がシステム開発に深く関与するという内製化文化がない先は、高速な機能改善や新技術・新サービスの登場にスピード感をもって対応していくことができていない。

副島論文は、生成 AI とりわけ大規模言語モデルの発展の歴史と、代表的なモデルの技術解説、現在の最新モデルやサービス群の紹介、実装技術のデモンストレーション付き解説を提供している。特に、企業での活用において鍵となる秘匿情報（企業内部情報や顧客情報）を生成 AI モデルで扱う手法を中心に実装例を紹介している。生成 AI を活用していくための方法を学ぶチュートリアルを意識して執筆されており、生成 AI 技術の民主化メリットを享受していくための水先案内（Pilot）となっている。

デジタルアイデンティティを巡る世界の潮流

柴田 健久｜PwC コンサルティング合同会社ディレクター
崎村 夏彦｜OpenID Foundation 理事長、PwC Japan グループ Digital Identity 顧問

柴田 健久
PwC コンサルティング合同会社ディレクター
地政学リスクや経済安全保障、各国の政策や規制、サイバー脅威を扱う Trust & Risk Consulting 部門に所属。
デジタルアイデンティティ技術を駆使した KYC、認証認可などが専門。
大手シンクタンクを経て PwC コンサルティング合同会社に入社。デジタルアイデンティティ技術をコアとする事業企画などを担当。

崎村 夏彦
PwC Japan グループ Digital Identity 顧問、OpenID Foundation Chairman
デジタルアイデンティティおよびプライバシー関連技術の国際標準化を専門としており、現在世界で 30 億人以上に使われている JWT、JWS、OAuth PKCE、OpenID Connect、FAPI、ISO/IEC 29100、ISO/IEC 29184 などの国際規格の著者・編者。

要約

デジタル社会は急速に進化し、金融業界においても、セキュリティの確保、プライバシーの保護、規制の遵守、そしてオープンバンキングの導入と企業間のデータ共有などが進んできた。このような中で、シームレスなユーザ体験を提供し、個々のユーザが自身の情報を管理できるようなデジタルアイデンティティの整備がますます重要となっている。

この論文では、オープンバンキングの動向に触れた上でデジタルアイデンティティが今後担う役割を説明したのち、金融業界に関係が深いと考えられる新たな技術であるデジタルアイデンティティウォレットを紹介する。そして最後に、デジタルアイデンティティ整備にあたり今後求められる相互運用性、政府を含む市場に求められる対応について考察する。

1．はじめに

現代社会は、デジタル技術の急速な発展により、かつてないスピードで変化している。API の普及で様々なサービスを組み合わせたサービスが生まれ、さらに時間の経過とともに生成された膨大なデータによって、新しいサービスが日々生まれている。ユーザも、様々な業種で自分のニーズや好みに合ったサービスが利用可能となり、ユーザ中心主義の視点で設計されたサービスを使用するデジタルライフを送ることができるようになってきた。

金融業界もこの例外ではなく、デジタル化・オープン化の波に直面している。オープンバンキングやオープンファイナンスの概念が広まり、金融機関は顧客データの共有と連携を通じて、より革新的で顧客中心の金融サービスを提供することが求められている。

この変革の中で、デジタルアイデンティティは、次世代のデジタル金融サービスを支える重要な役割を果たす。

本稿は、このデジタルアイデンティティに関する世界の動きについて紹介する。

2．デジタル社会とデジタルアイデンティティ管理の重要性

デジタル社会の進展に伴い、個人のアイデンティティ管理はより複雑かつ重要になっている。オンラインサービスの普及で便利となった一方、デジタル犯罪も多く発生しており、個人情報の保護とセキュリティの確保が喫緊の課題となっている。

デジタルアイデンティティは、個人に関するデジタル情報の集合体である。これには、氏名、住所、生年月日などの基本的な情報に加え、生体情報や行動データなど、あらゆる情報が含まれる。裏を返せば、デジタルアイデンティティを適切に管理することで、オンラインサービスへのアクセス制御、不正防止、個人情報の保護などを実現することができる。

金融業界においては、デジタルアイデンティティの管理は特に重要である。

オンラインバンキングはほぼすべての金融機関で導入されており、すでにデータの安全性と顧客のプライバシー保護が最優先事項となっているが、昨今はオープンバンキング、オープンファイナンスを導入する動きが各国で活発化しており、顧客データの共有と連携が進む中、さらに重要性が増してきている。

3．各国のオープン化に向けた取り組み状況

オープンバンキングについて、日本では 2017 年に改正銀行法が成立して API の提供が努力義務となった。その後も金融情報システムセンター（FISC）による API 接続チェックリスト[1]が策定されるなどで徐々にその活用が広がっているが、世界各国では、オープンバンキングの導入に向けた取り組みが進められている。

1:「API 接続チェックリスト〈2018 年 10 月版〉」一部改訂のお知らせ〈API 接続チェックリスト〉（2024-05-05 取得）

図1　各国のオープンバンキング状況

出所）筆者作成

ヨーロッパは早くからオープンバンキングに取り組んでおり、英国では2014年以降競争・市場庁（CMA）が主導して、オープンバンキングの標準APIの開発と導入、制度化が進められている。欧州連合（EU）でも2015年に制定された改正決済サービス指令（Payment Services Directive：PSD）2により、オープンバンキングの法的枠組みが整備された。

また、米国では包括的なオープンバンキング規制はないものの、市場主導型のオープン化が進んでいる。大手銀行やフィンテック企業が自主的にAPIを公開し、サードパーティーとの連携を進めている。2023年10月 に米消費者金融保護局（CFPB）が金融機関のAPIアクセスの無償化等の制度化を提案[2]。カナダでは、2024年中にオープンバンキングのフレームワーク（Consumer-Driven Banking Framework）を導入予定[3]と発表された。

アジアでも多くの国でオープン化が進んでいる。シンガポールでは、2015年に金融管理局（MAS）がフィンテック環境の整備を目的とした「FinTech & Innovation Group」を設立[4]し、オープンバンキングの推進に取り組んでいる。また、オーストラリアでは、2020年7月から段階的に実施されている[5]。

南米では、ブラジルがオープンバンキングの導入をリードしている。ブラジル中央銀行は2021年にオープンバンキング規制を施行し、金融セクターの競争力強化と金融包摂の促進を図っている。現在、5億口座以上が月間50億トランザクションをさばいているなど急速な成長を示している[6]。また、これに続く形でコロンビア、チリなどがオープンバンキングを推進している。

中東地域では、アラブ首長国連邦（UAE）、サウジアラビア、バーレーンなどが導入しており、金融セクターの革新と顧客サービスの向上を目指している。UAEでは特にドバイとアブダビの金融センターがオープンバンキングを推進[7]しており、サウジアラビアではOpen Banking Frameworkがすでに導入されている。

アフリカでは、ナイジェリアやケニアなどで進展が見られる。ナイジェリア中央銀行は2021年にオープンバンキング規制の枠組みを発表[8]し、金融包摂の促進を目指している。ケニアでは、モバイルマネーサービスが広く普及しており、オープンバンキングへの基盤が整っている。

2：CFPB Proposes Rule to Jumpstart Competition and Accelerate Shift to Open Banking< https://www.consumerfinance.gov/about-us/newsroom/cfpb-proposes-rule-to-jumpstart-competition-and-accelerate-shift-to-open-banking/>（2024-05-05取得）

3：Budget 2024: Canada's Consumer-Driven Banking Framework<https://www.canada.ca/en/department-finance/programs/financial-sector-policy/open-banking-implementation/budget-2024-canadas-framework-for-consumer-driven-banking.html>（2024-05-05取得）

4：MAS sets up new FinTech & Innovation Group<https://www.mas.gov.sg/news/media-releases/2015/mas-sets-up-new-fintech-and-innovation-group>（2024-05-05取得）

5：Australian Banking Association「Open Banking」<https://www.ausbanking.org.au/priorities/open-banking/#:~:text=Open%20banking%20gives%20you%20the,products%20or%20banks%20more%20easily.>（2024-05-05取得）

6：Banco Central do Brasil（2024），Pix Statistics < https://www.bcb.gov.br/en/financialstability/pixstatistics>（2024-05-15取得）

7：Arab Regional Fintech Working Group「Open Banking Regulatory Principles」（2021/3）<https://www.amf.org.ae/sites/default/files/publications/2021-12/open-banking-regulatory-principles.pdf>

8：REGULATORY FRAMEWORK FOR OPEN BANKING IN NIGERIA <https://www.cbn.gov.ng/out/2021/psmd/circular%20on%20the%20regulatory%20framework%20on%20open%20banking%20in%20nigeria.pdf>（2024-05-05取得）

4．次世代デジタル金融とデジタルアイデンティティの関係

デジタル化とオープンバンキングの進展により、次世代のデジタル金融サービスが登場しつつある。

オープン API を通じたデータの共有と連携は、革新的な金融サービスの開発を加速させている。

欧州委員会は PSD3 の提案に際して公開したデータで、EU における電子決済は、2017 年の 184.2 兆ユーロから、2021 年には 240 兆ユーロに増加したとしている[9]。

9：https://ec.europa.eu/commission/presscorner/detail/en/fs_23_3558

これらの次世代デジタル金融サービスを支えるのが、デジタルアイデンティティである。デジタルアイデンティティは、ユーザを一意に識別し、認証し、プライバシーデータを管理し、利活用するために必要な機能を持つ。これにより、ユーザは個人情報を安心して預けられるし、事業者はユーザの同意を得られた範囲でデータを分析し、ユーザのニーズに合ったサービスを提供することができる。

さらに、デジタルアイデンティティは、金融包摂の実現にも大きな役割を果たす。

デジタル化されたセキュアな本人確認・認証により、これまで金融サービスへのアクセスが限られていた層に対して、新たな金融サービスを提供することが可能になる。また、モバイルマネーサービスなどのデジタル金融サービスを安全かつ効率的に展開することができる。

次世代のデジタル金融サービスにおいては、セキュリティとコンプライアンスの確保が非常に重要である。特に、デジタルオンボーディングプロセスでは、金融サービス機関は身元確認を厳密に行うことが必要である。これにより、新規アカウント詐欺、アプリケーション詐欺、カード情報盗難、アカウントの乗っ取りなどから自身と顧客を守ることができる。また、銀行支店で発生する特定の種類の詐欺を防ぐためにも、デジタル身元証明の利用が有効である。

また、オープン API を通じたデータの共有と連携をする場合、プライバシー保護、セキュリティ確保とリスク管理をエコシステム全体で構築する必要がある。そのためには、金融機関、フィンテック企業、規制当局が協力し、データの適切な取り扱いとセキュリティ対策を確保することが不可欠である。そして、オープンバンキングにおけるデータ保護とセキュリティに関する明確なガイドラインを策定し、関係者がこれらのルールを遵守していることを監督する仕組みをつくることが必要だろう。

次世代のデジタル金融サービスにおいては、デジタルアイデンティティの管理が重要な役割を果たす。金融機関、規制当局、テクノロジー企業が協力し、デジタルアイデンティティの管理体制を強化しながら、革新的で包摂的なデジタル金融の未来を築いていくことが求められている。

5．次世代デジタル金融におけるデジタルアイデンティティの役割

これまで述べたように、次世代デジタル金融において、デジタルアイデンティティには大きな貢献が期待されている。まず、デジタル身元確認と呼ばれる技術を活用することで、オンラインでの本人確認が容易になり、金融サービスへのアクセスが大幅に向上する。これにより、これまで金融サービスを利用できなかった人々も、容易にサービスを利用できるようになる。また、複数のフィンテックサービスで同様の手続きをしなくてもよくなることで、ユーザはサービスごとに異なるアカウントを管理する必要がなくなり、利便性が飛躍的に向上する。これにより、金融サービスの実利用率が向上し、金融包摂の進展に寄与することが期待される。

しかし、これらの期待に応えるためには同時に、デジタルアイデンティティがいくつかの重要な役割を果たす必要がある。具体的には、個人のプライバシー保護と信頼性の向上、金融犯罪の防止とコンプライアンスの強化、そして分散型アーキテクチャによる相互運用性の実現である。以下では、これらの役割について詳しく説明する。

5.1 個人のプライバシー保護と信頼性の向上

デジタルアイデンティティが果たすべき最も重要な役割の一つは、個人のプライバシー保護と信頼性の向上である。プライバシーについては、日本の個人情報保護法の他、EU の EU 一般データ保護規則（General Data Protection Regulation) 等、各国・地域で法整備が進んでいる。米国でも現在、米国プライバシー権法（American Privacy Rights Act）が審議中である。これらの法制度は、消費者のプライバシーを守るために、事業者に対して消費者データの収集、使用、販売に関する高い透明性と、データプライバシーとセキュリティの管理監督を義務付けている。

そして、その実現には個人、その個人から個人データを収集する主体、個人データを処理する主体の役割と責任を明確にし、適切に運用することが必要である。これらが実現できれば、信頼性が大幅に向上する。

他に個人情報の開示についても、個人が状況に応じて選択的に開示することができるような技術の標準化が進められているが、このような技術が普及すれば、個人は自分の情報を自分の意思で管理できるようになり、プライバシーと利便性の両立に寄与し得る。

5.2 金融犯罪の防止とコンプライアンスの強化

デジタルアイデンティティのもう一つの重要な役割は、金融犯罪の防止とコンプライアンスの強化である。

そのためにはまず、信頼できる証拠に基づく厳格な本人確認プロセスの実施と、その実施記録の保存をすることが重要である。このことにより、なりすましや不正アカウントの作成の抑止、不正行為の調査や監査が可能となる。

さらに、マネーロンダリングや不正送金などのリスクの高い取引にデジタルアイデンティティをキーとする行動分析技術を適用すれば、利便性とのバラ

ンスを保ちながら不正リスクの低減や規制当局からの要件遵守に寄与するだろう。

日本ではすでに、金融機関は eKYC やマネーロンダリング対策（AML）をより効果的に行うことができるようになり、効率化が進められてきた。2018年の「犯罪による収益の移転防止に関する法律施行規則」の改正でオンラインでの本人確認が認められ、PC やスマートフォンで撮影した本人確認書類の送付や公的個人認証サービス（JPKI）の署名用電子証明書（マイナンバーカードに搭載されている署名用電子証明書）の利用が可能となっている。

ただし現在、各国ではさらにこの見直しが進もうとしている。

例えばデジタル庁が公表する「DS500 行政手続におけるオンラインによる本人確認の手法に関するガイドライン」（DS500）のベースの一つとなった米国 NIST の「SP 800-63 Revision 3 Digital Identity Guidelines」は、Revision 4 として更改される動きがある。ここには、リモートで厳密な身元確認を実施する場合、訓練されたオペレータによる監視下であること[10]などが記載されているが、現在日本はこれを必須としていない。

他にも、英国でも 2024 年 1 月に身元確認の基準となるガイド「How to prove and verify someone's identity Good Practice Guide 」[11]が更新された。また、AML にあたってデジタルアイデンティティの技術をどのように活用できるか、具体的なガイダンス作成が検討されている。

今後、国境を越えたデータのやり取りを行うことを視野に入れた場合、これらのグローバル規制との乖離はできるだけ発生しないようにすべきである。2023 年 6 月には犯罪対策閣僚会議より「国民を詐欺から守るための総合対策」[12]がとりまとめられ、オンラインの本人確認手法はマイナンバーカードの公的個人認証に原則として一本化し、運転免許証や顔写真のない本人確認書類等は廃止する方針が示された。また、上記の DS500 も現在、グローバルとの乖離があった場合、どう向き合うのかデジタル庁にて見直しの議論が進められている。当然、民間サービスへの波及も考えられることから注目が必要である。

5.3 シームレスなデジタルライフを実現する仕組みと相互運用性の実現

次世代デジタル金融においては、様々な金融サービスが相互に連携し、シームレスなユーザ体験を提供することが求められる。しかし現状では、多くの金融機関が独自のデジタルアイデンティティシステムを構築しているためサイロ化が進んでおり、サービス間の連携が困難になっている。

この問題を解決するためには、接続が容易なデジタルアイデンティティアーキテクチャを採用し、相互運用性を実現することが不可欠である。これができれば、様々なシーンで検証可能なアイデンティティの発行や管理を行うことができる。これにより、異なる金融サービス間でもアイデンティティの相互運用性を確保することができるであろう。

現在、分散型アイデンティティの標準技術として、Verifiable Credential や OpenID Connect の拡張仕様の検討が活発となっている。これらの標準技術を採用することで、金融サービス間でのアイデンティティの相互運用性を実

10：National Institute of Standards and Technology:NIST (2023), NIST SP 800-63 Digital Identity Guidelines SP 800-63A 5.5.3 In-person Proofing Requirements <https://pages.nist.gov/800-63-4/sp800-63a.html#vip> (2024-05-05 取得)

11：イギリス政府 (2023), Guidance How to prove and verify someone's identity <https://www.gov.uk/government/publications/identity-proofing-and-verification-of-an-individual> (2024-05-05 取得)

12：犯罪対策閣僚会議「国民を詐欺から守るための総合対策」（総務省 < https://www.soumu.go.jp/main_content/000953287.pdf>）(2026-07-04 取得)

現し、ユーザの利便性を大幅に向上させることが可能となるだろう。

本節では、次世代デジタル金融におけるデジタルアイデンティティの役割について紹介した。デジタルアイデンティティは、個人のプライバシーを保護しつつ、金融サービスのアクセシビリティと利便性を向上させるために不可欠である。また、金融犯罪を防止し、コンプライアンスを強化するためにも、デジタルアイデンティティのセキュリティ要件が重要である。

さらに、異なる金融サービス間でのアイデンティティの相互運用性を確保するためには、標準化が不可欠である。

次節では、これらの点についての最新動向を紹介する。

6．デジタルアイデンティティの技術動向

6.1 Verifiable Credentialの必要性

サービス登録に必要なアイデンティティ情報は多岐にわたる。具体的に3つの事例を取り上げて確認してみよう。

法人口座開設の例

- 履歴事項全部証明書（商業登記簿謄本）：法人の基本情報が記載されている公的な書類。法務局で取得可能。
- 法人の印鑑証明書：法人が登録している印鑑の証明書。法務局で取得可能。
- 代表者の本人確認書類：代表者の身分を証明するための書類（運転免許証、パスポートなど）。
- 法人番号が確認できる書類：法人番号を記載した書類。法人番号通知書などが該当。
- 事業実態を証明する資料：会社案内、ホームページのプリントアウト、契約書のコピーなど、事業が実際に存在していることを示す資料。
- 委任状：代表者以外の人が口座開設の手続きを行う場合に必要な書類。

携帯電話契約の例

- 運転免許証
- マイナンバーカード（個人番号カード）
- パスポート
- 在留カード
- 身体障害者手帳、精神障害者保健福祉手帳、療育手帳
- 預金通帳＋お届け印（クレジットカードやキャッシュカードがない場合）

大学院入学の例

- 卒業証明書ないし卒業見込み証明書
- 成績証明書
- 推薦書

- 健康診断書

これらは従来、紙やプラスチック板の目視やコピー、またはそれに準ずる手段で確認されることが多かった。
しかし、偽造技術の進展とともにこの検証が非常に困難になってきており、一人数百万円オーダーの被害が相次ぐようになってきている。

これをデジタル化して容易に検証が可能にしようというのが検証可能クレデンシャル（Verifiable Credential：VC）の取り組みである。これらがデジタルで自動検証できるようになることで、事故が減るだけでなく、経済効率性も大幅に向上し、GDP の増加にも寄与すると考えられる。そのためには、こうした証明書類がすべからず検証可能資格証明書として提供されるようになることが肝要である。

VC は証明対象となるデータに、証明書発行者の電子署名ないしは電子シールがついたものだ。現在主流のフォーマットには CBOR ベースの mdoc（ISO/IEC 18013-5 で定義）と、SD-JWT VC（IETF で定義）があるが、それ以外にも多種多様なものが存在する。日本政府がマイナンバーカード関連で利用している X.509 も広い意味ではこの中に含まれる。Open Wallet Foundation は 2024 年 5 月 8 日時点で 16 種類をリストしている[13]。

これを対象者 (subject) が操る（通常はスマホ上で稼働する）ウォレットと呼ばれるソフトウェアに発行、格納する。発行に使われるプロトコルとして、例えば、EU が制定中のウォレット関連の規格である EU Digital Identity Architecture and Reference Framework 1.4（ARF 1.4）[14] では OpenID for Verifiable Credential Issuance（OpenID4VCI）が指定されている。
こうして保管された VC は、対象者のウォレット操作によって「提示（presentation）」と呼ばれるプロセスを通じてその利用者（RP、検証者）に検証可能提示（Verifiable Presentation：VP）として提供される。VP には対象者がウォレット上で生成した鍵による署名がついている。ARF 1.4 では、この提示のプロトコルとしては、近接提示は ISO/IEC 18013-5 が、遠隔提示には、OpenID for Verifiable Presentation が指定されている。

検証者はこれを受け取り、VC にかかっている署名を確認し、署名者の正当性を確認することによって、証明書記載事項が正しいことを確認できる。
また、提出者と証明書に記載されている人が等しいことは（ここでは提出者同一性とよぶ）、証明書に含まれる
A）提出者の生体情報
B）提出者の鍵情報（ウォレットソフトウェアが生成したもの）
などによって確認可能である。

生体情報は主に提示を受けるのが「人」であるケースを想定している。VC

13：OWF(2024) credential-format-comparison-sig/data/Credential-Format/ https://github.com/openwallet-foundation/credential-format-comparison-sig/tree/main/data/Credential-Format

14：European Commission (2024), EU Digital Identity Architecture Reference Framework 1.4 < https://github.com/eu-digital-identity-wallet/eudi-doc-architecture-and-reference-framework/blob/main/docs/arf.md> (2024-07-04 取得)

に入っている顔写真と提示者を目視確認するような場合だ。専用ハードウェアが利用できる場合、この目視はカメラ画像による確認で代替できるかもしれない。

鍵情報の場合は、VC を VP として提示する時に、その鍵による署名が付されることによって示される。

提出者の鍵情報を VC に含めることによって提出者同一性を確保する場合には、VC に含む鍵が本当に想定する個人や法人のものであるか、発行先となるウォレットが正しいものであるかを確かめる必要がある。この確認の厳密さによって、欧州では 2 段階のものが設定されている。

6.2 従来のモデル（アイデンティティ・フェデレーション）との違い

こうした属性連携のデジタルな仕組みで現在主流であるのは、OpenID Connect などのアイデンティティ連携（Identity Federation）の方式である。OpenID Connect でも、VC における発行者にあたる Claims Provider (CP) というものが存在する。ここが署名付きの属性証明を出し、これを OpenID Provider (OP) と呼ばれる上記のウォレットにあたるソフトウェア（ただし、通常これはクラウド上のシェアードサービスとして実装される）を通じて提供される形だ。トポロジー的にはほぼ同じものになるが、実態上はいくつかの大きな差がある。

① OpenID Connect の場合、検証者が受け取るデータには OP の署名（対象者の署名とは異なる）がついている。したがって、個人の署名の代替としては使えない。一方で、EU Digital Identity Wallet（EUDIW）では署名サービスを提供することになっている。

② ウォレットは殆どの場合ユーザは一人であるのに対して、OP は大量の人でシェアする。これは、受け取り手による個人の識別性を下げる（k- 匿名性の確保）という意味では望ましいが、多量のユーザをプロバイダが観測して広告に使うなどという観点からはプライバシー的に望ましくないとされる（ただし、EUDIW がそうしているように、法規制すれば良い問題ではある。逆に、これができないようにすることによって EUDIW にはビジネスモデルがなくなり、民間レベルでの運用が難しいということも指摘されている。）。

③ OP はネットワーク上で常にオンラインであることが期待され、検証者は随時最新の情報を得ることができるが、ウォレットはオフラインであることが期待され、ユーザが行動を起こした時にしか情報が提供されない。

④ CP からの情報を OP が中継する方式は標準で定義されているものの、実際にはほぼ使われておらず、各 CP が OP として振る舞って直接属性提供する場合が殆ど。一方、EUDIW では直接提供のパスは用意されておらず、必ずウォレット経由での提供となるため、中継機能が必ず活用されるようになることが期待される。また、これによってユーザからすると一括管理がしやすくなるというメリットがある。

⑤ OP は CP から事前に発行を受けた属性証明はその一部を選択的に提供す

ることはできない。選択的に提供しようとすると、都度 CP から必要なものだけを取得することになり、これは CP も 24 時間 365 日運用が必要であるということになり、運用負荷が重い。

⑥ 複数の CP からの情報を連携しようとするとそれだけ取得時間がかかることになり、あらかじめ取得しておいてまとめて提示できるウォレットモデルに対して使用体験が悪くなる。

また、一般の方がウォレットを初めて知ったときに受ける印象として、

① ウォレットはユーザデバイス上で稼働するのでユーザのコントロール下にあり、プライバシー的に有利

② 個人の情報がサーバに蓄積されないので、抜かれることがなくなり安全

といった点も指摘されるが、実はこれらは必ずしも正しくない。例えば、対象者に関する情報をウォレットがバックエンドのサーバに送って利用するということは十分に考えられる。また、属性情報をオンラインで集積している問題は、元となる情報はもとより VC の発行者に溜まっているのであり、そこを攻撃すれば OP を攻撃するのと同様かそれ以上の効果がある。また、ある個人を攻撃するという観点で言えば、ウォレットに溜まった情報を詐欺サイトなどを通じて一気に抜く方が効率が良い。そのため、上記のメリットを享受しようとすると、追加で条件が必要になる。EU のデジタルアイデンティティフレームワークでは以下のような条件が課されている。

① ウォレットのアプリ部分はオープンソースで提供されなければならない（EU DIF 5a-3）。

② ウォレットの提供者に、プライバシー保護技術と不可観測性を実装し、提供者がユーザの行った取引の詳細を見ることができないようにしなければならない。また、その提供者は審査機関によって審査・認定を受けなければならない。

③ VC を利用する RP/ 検証者はその正当性を示すために、ウォレットに対して運用者情報を含めてクライアント認証を実施できなければならない。

④ PID の発行を受けるウォレットは認定されたものでなければならず、Wallet Instance Attestation というものを使ってそのインスタンスのクライアント認証が可能である必要がある（なお、こうしたウォレットは特定の認可されたスマートフォンの上でしか稼働しない模様。2024 年 4 月時点 4 種類のみとされる[15]。）。

ウォレットのオープンソース化ということに関しては、すでに EU ではオープンソースのリファレンス実装が公開されており、各国はこれを元に実装を行っていく方針となっている。また、このソースコードは Open Wallet Foundation にとりこまれ、そこでメンテナンスが行われていく見込みである。ただし、これにかかる費用を誰がどのように負担していくのかというのはまだ検討課題として積み残されている。

また、そのウォレットの提供者の運用に関しては、上記のとおり厳しく規制

15：Bradley, (2024) Split Key ECDSA and ARKG for Wallet Proof of Possession, IIW38 Notes https://docs.google.com/document/d/1DjOS1MOjJtt0BporEHXrcfHnfbfJmif9uXhHEhGIC1s/edit（2024-05-15 取得）

がなされる。

加えて、RP の正当性およびその要求する属性が必要最低限のものになっているかの確認が第三者によって行われ、それが明示されることが望ましいと思われる。

この他、EUDIW には以下のような機能が期待されている。

- EUDIW は、EU 市民と居住者が自分の個人データとアイデンティティ関連情報を安全、ユーザフレンドリー、かつ透明性の高い方法で管理、保存、共有できるように設計されたツールである。
- このウォレットにより、ユーザは自分の個人データと属性の電子的証明を安全に要求、取得、選択、組み合わせ、保存、削除、共有、提示することができ、データを信頼当事者に選択的に開示することができる。
- ウォレットを通じて実行されたすべての取引を追跡し、信頼当事者に個人データの即時消去を要求し、個人データの疑わしい要求を管轄の国内データ保護機関に報告することもできる。
- ゼロ知識証明などのプライバシー保護技術を統合し、基礎となるデータを明らかにすることなく、ユーザの識別データに基づいて記述を検証することができる（ただし、ARF 1.4 にはゼロ知識証明は入っておらず、選択的属性開示だけが入っている。）。
- ユーザには、EUDIW の設計に組み込まれた共通のダッシュボード機能が提供され、自分の個人データに対する透明性、プライバシー、コントロールを持つことができる。
- データの取り扱いに関する目的の制限、収集するデータの最小化、設計段階からデータ保護の観点を含め、利活用と保護の設定を可能とする場合は保護をデフォルト設定する。

全体として、このウォレットは、民主主義社会、基本的権利、法的保護措置を守ることを目的として、透明性が高く、ユーザが管理でき、相互運用性のあるように設計されることが期待されている。

また、プライバシー面に関しては、電子識別と認証に最高レベルのデータ保護とセキュリティを課している点を強調しており、ユーザが EUDIW の使用とデータを完全にコントロールできること、そして、そのようなウォレットの使用は任意であり、それを使用しない人に対して公的または私的サービスへのアクセスを制限しないことを保証するものになっている。

7．今後の展望と課題

以上、次世代デジタル金融におけるデジタルアイデンティティの重要性と、その実現に向けた技術動向について述べてきた。

ユーザが自身のアイデンティティをコントロールできる分散型のアーキテクチャとして、DID（分散型識別子）や VC などの具体的な実現手段が成熟し、関連するフレームワークや規制が進歩することは、異なるサービス間でのアイデンティティの相互運用性の実現を後押しし、デジタルアイデンティティがパーソナライズされた金融サービスの提供や、異なる金融サービス間でデータを連携させ、より最適化された金融サービスの提供が可能になる。

最後に、特に注目すべき 2 つの動きとその課題について紹介する。

7.1 ウォレットの普及と官民の役割分担

前述のように eIDAS2.0 が承認され、EU では EUDIW の普及と社会実装が進められることが予想される。

また、日本では 2023 年来、スマートフォンでマイナンバーカードが提示できるようになってきている。同じく 2023 年には「マイナンバーカード機能等のスマートフォンへの搭載に係る実証事業（技術検証・要件検討）」の調達公募などもあり、政府によるウォレットの提供に向けた検討が進められていると考えられている。

そして、このウォレットに関する議論が国内外で活発に行われている。

ウォレットを使用することで、ユーザは自身の個人データを安全に管理し、必要な場面で適切に提示することができるようになることが期待されている。また、政府発行のアイデンティティ情報でユーザオンボーディングしたり、民間事業者の提供する個人データを政府の提供するウォレットに格納したりすることで、利便性が高まる可能性がある。

民間事業者にとっても、本人確認のコストを下げるなどのメリットを享受できる可能性もある。

ただし、政府が提供するアプリケーションに個人データを格納し、様々なシーンで利用できるようになることは、政府が国民がいつどんなオンラインサービスを使っているのかなど、プライバシー情報へアクセスできる可能性があることを意味する。

この点について、日本で議論になっている他、同じように EU でも多くの意見が寄せられたとされており、今後、管理・監督方法や透明性確保の方法などの議論が必要な状況となっている。

7.2 国際的な相互運用性の確保

オープンバンキング、ウォレットなどの普及には大きな期待がかかる反面、参加者の間で信頼性の高い異業種・国際間の連携をシームレスに実現するには、相互運用性の確保はますます重要な課題となるだろう。

こうした中 2024 年 5 月、デジタル庁が EU とデジタルアイデンティティに関する協力覚書を交わしたとの発表があった[16]。将来的なアイデンティティマネジメント体系の相互承認に向けた課題についても検討を始めるとされている。

デジタルアイデンティティ技術の発展は、次世代デジタル金融におけるビジ

16：デジタル庁（2024）, 河野デジタル大臣はブルトン欧州委員と会談を行い、EU とデジタル・アイデンティティに関する協力覚書を交わしました <https://www.digital.go.jp/news/eea22370-19d8-4a1a-ae92-89e28476f9a1>（2024-05-05 取得）

ネスの加速とセキュリティの向上に大きく貢献することが期待される。そのために技術や相互運用性の確保のための議論は開始されており、政府側の動きが活発である。

金融機関が、デジタルアイデンティティを核とした新たなビジネスモデルの創出に積極的に取り組むのであれば、これらの動向は注視すべきであり、必要に応じて参加していくべきであろう。

「欧州デジタル ID 枠組み規則」制定の経緯と欧州デジタル ID ウォレットの共通仕様

― EUDIW Architecture framework v1.4からみる技術仕様 ―

中山 靖司｜SBI 金融経済研究所 主任研究員　SBI 大学院大学 客員教授
NPO 法人金融 IT 協会 理事

中山 靖司
SBI 金融経済研究所 主任研究員
SBI 大学院大学 客員教授
NPO 法人金融 IT 協会 理事
1964 年生まれ。東京工業大学（院）経営工学修士。1988 年日本銀行入行。主に情報セキュリティや電子決済関係の実務および調査・研究に従事し、1996 年日銀における CBDC 研究の先駆け「日銀－ NTT 方式」電子通貨を研究・開発。東京大学先端経済工学研究センター（現在先端科学技術研究センターに吸収）助教授、FISC 調査部長等を歴任後、日銀金融高度化センターで「IT（AI）を活用した金融の高度化に関する WS」（全 10 回）を企画し座長を務める。特許「発行機関分離型番号登録式電子現金方法および利用者装置」他。

要約

本稿は、2024 年 5 月に発効された「欧州デジタル ID 枠組み規則」制定の経緯を紹介するとともに、その規則に基づき EU 市民に提供することが義務付けられた欧州デジタル ID ウォレット（EUDIW）のユースケースや技術仕様（設計原則、エコシステム、個人識別データの概要等）について解説したものである。EUDIW は、公共および民間のデジタルサービスを受ける際に本人確認の手段として利用できるアプリであり、ユーザーが自身の ID データや属性情報を安全に保存・管理できる等の特徴がある[1]。一方、日本でも、マイナンバーカードに健康保険証や運転免許証の情報を紐づける取組みや、マイナンバーカード自体をスマートフォンに載せるための法整備等が進んでいるが、EUDIW のような、本人認証にかかる属性情報を管理するデジタル ID ウォレットの発想からデザインされたものではない。そのため、EUDIW を参考に、将来を見据えてグランドデザインを見直すことが必要ではないだろうか。

1．はじめに

「欧州デジタル ID 枠組みの構築に関する規則＜（EU）第 910/2014 号を改正する欧州議会および理事会規則＞」（欧州デジタル ID 枠組み規則）が欧州議会および EU 理事会の双方によって承認され、2024 年 5 月 20 日に発効された[2]。

本規則は、2014 年に制定された「域内市場における電子取引のための電子

1：一般に、「ウォレット」とは主にスマートフォンやタブレットなどのデバイスに搭載される、財布の役割を提供するものであり、支払い機能に加えて、デジタル資産・ID 属性情報などの管理もできるとされる。EUDIW は、ID 属性情報管理機能の面からウォレットを活用するものであり、デジタル資産管理や支払い機能については触れられていない。

2：本規則は、2021 年 6 月に欧州委員会によって提案され、欧州議会、EU 理事会において検討作業が進められていたが、2023 年 6 月、提案の主要な要素についての暫定的な政治的合意に達し、同年 11 月、三者協議で最終合意した。これを受けて、同規制案は 2024 年 2 月に欧州議会で、4 月に EU 理事会で採決が行われ、2024 年 5 月に発効された。これで EU における正式な手続きが一区切りついたことになり、各加盟国は 2026 年までに対応を迫られることになった。

的本人確認およびトラストサービスに関する規則」（eIDAS 規則）[3] を改正するもので、EU 域内で公共サービスを安全に利用し、オンライン上および国境を越えて取引を行うための基礎を築くものである。特に、欧州デジタル ID ウォレット（以下、「EUDIW」と呼ぶ）を通じて、すべての EU 市民、居住者、企業が利用できる欧州デジタル ID の枠組みについて定めているところが注目される。

今後、技術仕様と認証の概要を定めた複数の実施法が採択され、加盟国は 24 カ月以内に、EUDIW を市民に提供する義務を負うことになる。規則承認後 6 ～ 12 カ月で定められた期限[4] までに採択されるこれらの法律は、規則の制定プロセスと並行して検討作業が進められている、技術的な共通仕様や要件を定める「EU デジタル ID ツールボックス」と整合するものとなることが求められており、これによって欧州全体でウォレットが統一的に実装されることが保証されることになった。

図表1　EUDIWに関する規則制定の流れ

2014/7/23	「域内市場における電子取引のための電子的本人確認およびトラストサービスに関する規則」（EU）第 910/2014 号（「eIDAS 規則」）制定
2020/10/1-2	**欧州理事会（European Council）** 欧州委員会に対し、相互運用可能な電子署名を含む、安全な公的電子 ID のための EU 全体の枠組みを提案し、オンライン上の ID やデータを管理できるようにするとともに、公的、私的、国境を越えたデジタルサービスへのアクセスを可能にするよう求めた。
2021/3/9	**欧州委員会（European Commission）** 通達「2030 年デジタル・コンパス:デジタルの 10 年に向けた欧州の道」[5] において、2030 年までに、欧州連合（EU）とその市民が、信頼され、ユーザーが管理できるアイデンティティを広く普及させ、各ユーザーがオンライン上でのやり取りや存在を自分で管理できるようにする、という目標を掲げる。
2021/6/3	**欧州委員会「欧州デジタル ID 枠組みの構築に関する規則**＜（EU）第 910/2014 号を改正する欧州議会および理事会規則＞」（「欧州デジタル ID 枠組み規則：eIDAS 規則改正案」）[6] を提出。
2022/12/14	**欧州議会（European Parliament）および EU 理事会（Council of European Union）**「デジタル 10 年政策プログラム 2030 の策定」[7] で、2030 年までに、信頼され、自発的で、ユーザーが管理するデジタル ID を広く普及させることを目的とした欧州連合の枠組みの目的とデジタル目標を定める。
2023/1/23	**欧州議会、欧州理事会および欧州委員会** 「デジタルの 10 年に向けたデジタルの権利と原則に関する欧州宣言」[8] で、EU に住むすべての人が、データ漏洩や個人情報の盗難や改ざんなどのサイバーセキュリティのリスクやサイバー犯罪から保護された、幅広いオンラインおよびオフラインのサービスへのアクセスを可能にする、安全かつ信頼できるデジタル ID を提供することを宣言した。また、すべての人が使用方法や共有先を自ら管理する等によって、個人データの保護を受ける権利があるとした。

3：「eIDAS 規則」の目的は、公共サービスを利用するための政府電子 ID（「eID」）の国境を越えた承認を可能にし、従来の同等の紙ベースのプロセスと同じ法的地位で国境を越えて承認されるトラストサービスのための連邦市場を確立すること。

4：期限は、欧州デジタル ID 枠組み規則の中で、内容により 2024 年 11 月 21 日ないし 2025 年 5 月 21 日までとされている。

5：COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS 2030 Digital Compass: the European way for the Digital Decade COM/2021/118 final https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0118

6：EU における立法手続きでは、欧州委員会（European Commission：EU の行政執行機関）が発議権を持ち、EU 理事会（Council of European Union：EU の立法機関）と欧州議会（European Parliament：EU 市民の代表）に法案を提出し、双方で可決されれば、正式に法案成立となり、官報掲載後 20 日後に発効する。同規則に関する具体的な立法手続きの流れについては、以下の URL を参照。
https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=celex:52021PC0281#2023-12-07_APR_AGRPROV_CONSIL_byEP_CMT

7：Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030 (Official Journal of the European Union, L323, 19.12.2022, p.4)

8：European Declaration on Digital Rights and Principles for the Digital Decade (Official Journal of the European Union, C23, 23.1.2023, p.1)

2023/2/10	European Digital Identity Wallet Architecture and Reference Framework（ARF）の初版公開[9]
2023/5 月	ARF を基にした4つの大規模パイロット・プロジェクトの開始
2023/6/29	**欧州委員会**　欧州議会および EU 理事会と、**「欧州デジタル ID 枠組み規則」**の主要な要素について暫定的な政治的合意に達する。
2023/11/8	**三者協議**（欧州委員会、欧州議会、EU 理事会）で最終合意
2024/2/29	**欧州議会**で最終承認
2024/3/26	**EU 理事会**　第 4016 回会合で採択＜ ST 8552 2024 ＞
2024/4/11	**欧州議会議長**および **EU 理事会議長**による**署名**[10]**(EU)第 2024/1183 号**
2024/4/30	**EU 官報に掲載**
2024/5/20	**発効**（EU 官報掲載から 20 日後）　同規則は 2026 年までに完全に施行
2024/5/20	European Digital Identity Wallet Architecture and Reference Framework（ARF）の 1.4 版公開 [11]

出所）　筆者作成

2.「欧州デジタルID枠組み規則」の制定の経緯

2.1 背景

従来の eIDAS 規則では、EU 加盟国は任意で国内の eID（オンラインサービスにおいて利用できるデジタル ID）制度を届け出ることができ、他の加盟国はそれを承認する義務を負っていた（義務化は 2018 年）。しかし、加盟国が eID を整備すること自体は任意であり義務ではなかった。また、様々な eID システムを接続する上部構造を構築することで、相互運用性を確保しようとしていたが、各国間の整合がとれないという技術的な問題が発生しやすく、民間のデジタルサービスへの拡大も妨げられていた。

こうした中、2020 年 10 月、EU 加盟国の首脳らをメンバーとする EU の政治的最高意思決定機関である欧州理事会は、欧州委員会に対し、相互運用可能な電子署名を含む、安全な公的電子 ID のための EU 全体の枠組みを提案し、公的、私的、国境を越えたデジタルサービスへのアクセスを可能にするよう求めた。

これを受け、欧州委員会が実態をまとめたところ、2018 年 9 月に eIDAS 規則の eID 部分が発効して以来、少なくとも 1 つの eID スキームを通知している加盟国は 27 カ国中わずか 14 カ国と約半分に過ぎず（2021 年現在）、その結果、国境を越えて信頼できる安全な eID スキームにアクセスできるのは、EU 居住者の 59% にとどまっていた。また、現在のユーザーニーズに応えられるよう、完全にモバイル対応まで済ませている eID スキームは 7 つしかなかった。

そこで、2030 年までに、少なくとも 80％の市民が、主要な公共サービスを利用する際にデジタル ID ソリューションを使用できるようにするため、現

9：The European Digital Identity Wallet Architecture and Reference Framework v1.0
https://digital-strategy.ec.europa.eu/en/library/european-digital-identity-wallet-architecture-and-reference-framework

10：Document 32024R1183, Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing the European Digital Identity Framework
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1183

11：European Digital Identity Wallet Architecture and Reference Framework v1.4.0
https://eu-digital-identity-wallet.github.io/eudi-doc-architecture-and-reference-framework/1.4.0/arf/#314-qualified-and-non-qualified-electronic-attestation-of-attributes-schema-providers

行の枠組みを改訂した欧州デジタル ID を新たに提供することを提案した。eIDAS 規則では提供が任意だったことから普及が十分でなかったこともあり、新しい規則では、加盟国に対し、国のデジタル ID を他の個人属性（運転免許証、卒業証明書、銀行口座など）の証明とリンクできるデジタルウォレットを市民や企業に提供することが義務付けられるものとなった。

2.2 EUデジタルIDウォレット（EUDIW）の特徴

EUDIW は、アプリの形をした ID や属性情報を管理する個人用デジタル財布であり、市民がデジタル形式で自分自身を識別し、ID データや公文書を保存・管理することを可能にする。その特徴は、① EU デジタル ID を利用したい EU 市民、居住者、企業は誰でも利用できること[12]、② EU 全域で公共および民間のデジタルサービスへのアクセスをユーザーに提供する際の本人確認手段として使用されること、③利用者が第三者と共有する ID、データ、証明書を選択し、追跡できるようにするとともに、共有する必要のないものは共有されないなど、ユーザー自らが EUDIW を管理できること[13] である。

ユースケースとしては、例えば、運転免許証、医療処方箋、教育資格などが含まれるとされるが、新しい EUDIW により、モバイルデバイスのボタンをクリックするだけで、身元を証明したり、電子文書を共有したりできるようになり、欧州のすべての人々は、欧州全域で利用可能な自国のデジタル ID を使ってオンラインサービスにアクセスできるようになる。その期待されるメリットは以下のとおり。

（市民と企業）
1. **ユーザーのコントロール**：市民は、自分のアイデンティティやデータのどの側面を第三者と共有するかを選択する権限を持ち、個人情報のプライバシーと管理を保証する。
2. **広範なユーザビリティ**：EU 全域で、公共および民間のデジタルサービスへのアクセスが可能となり、オンラインでのやり取りがよりシームレスで効率的になる。
3. **透明性と安全性**：オープンソースライセンスを採用し、透明性と安全性を確保する。誤用や違法な追跡を防ぐための対策が講じられ、データが安全に取り扱われる。
4. **使いやすさ**：ユーザーフレンドリーなインターフェースを提供し、個人が簡単にデジタル ID を管理し、サービスにアクセスできる。市民であれば誰もが電子署名を無料で使うことができる。
5. **スムーズな移行**：市民は、各国のスキームを使ってウォレットに情報を登録することができ、デジタル ID 管理へのスムーズな移行が保証される。

（政府）
1. **デジタルサービスへのアクセス向上**：ウォレットは本人確認のプロセスを合理化し、市民がオンラインで政府サービスにアクセスしやすくし、利用率を高めることができる。
2. **詐欺防止の強化**：安全で検証可能な ID 手段を提供することにより、政府サービスに関する ID 窃盗や関連詐欺を減らすことができる。
3. **セキュリティの向上**：市民データの全体的なセキュリティが強化され、侵害リスクが軽減される。

（デジタルサービスの提供事業者）
1. **セキュリティとプライバシーの向上**：ウォレットは、従来の認証方法の責任に関連するリスクを軽減することができる。
2. **認証コストの削減**：ウォレットは、本人確認プロセスを簡素化し自動化することで、本人確認プロセスに関連するコストを削減することができる。
3. **競合する大手プラットフォームへの依存回避**：サービス提供事業者は、取得したユーザー・データの利用が不透明な ID サービスへの依存度を下げなければならなくなる。

12：EUDIW の対象ユーザーは加盟国の国内法で定められることになる。なお、欧州デジタル ID 枠組み規則では、加盟国に対し EUDIW を提供する義務を課しているが、一方で、これを保有し使用することを義務付けるものではない。

13：個人データ処理は、一般データ保護規則（GDPR）に完全に準拠して実施されることが求められた。

（社会）
1. **オンライン取引の増加**：認証がより簡単で安全なため、人々はオンライン取引をより行う傾向が強まる可能性がある。
2. **新たなビジネスチャンス**：アイデンティティ・ウォレットの採用はイノベーションを促進し、新しいサービスや製品に繋がる可能性がある。
3. **リソースの再配分**：これまで手作業の検証プロセスに費やしていたリソースを、より生産性の高い用途に振り向けることができる。
4. **経済成長**：オンライン取引の導入拡大、新たなビジネスチャンス、資源配分の改善は、全体として経済の安定と成長に貢献する。

2.3「EUデジタルIDツールボックス」の開発について

多くの加盟国は、属性およびクレデンシャル（認証情報）の統合のために、デジタルウォレットや国家間のトラストフレームワークを含む国家デジタル ID システムを展開または開発している。しかしながら、各国が独自に異なるソリューションを開発することは、規格の相違による分断や障壁を生み、欧州単一市場の恩恵を奪うことになるとして、新たな規則では加盟国に対し、お互いのソリューションを認め合うだけでなく、共通の技術標準に基づいて構築されたスキームの下でデジタルウォレットを発行するよう求めている。そのため、欧州委員会では、本規則を提案するのに合わせ、ウォレットの技術仕様等を定義する「EU デジタル ID ツールボックス」の開発を並行して進めるよう各加盟国に対する勧告を付していた[14]。

EU デジタル ID ツールボックスは、技術アーキテクチャおよび参照フレームワーク、一連の共通の規格および技術仕様、ガイドラインとベストプラクティスからなるもので、加盟国の専門家が、関係官民団体と緊密に連携して開発にあたっている。また、EUDIW を活用する多数の大規模なパイロットテストにおいて、プロトタイプを設計するための基盤として活用することで、改善され精緻化されていくこととなる。

EUDIW の技術アーキテクチャおよび参照フレームワークとしては、Architecture and Reference Framework（以下 EUDIW ARF）の 1.4 版が、欧州デジタル ID 枠組み規則の法文に基づいた最新版（本稿執筆時）として、2024 年 5 月 20 日に公開された[15]。EUDIW ARF1.4 版は説明的な主文書のほか、6 つの付属文書から構成され、付属文書のうち、「付属文書 2：ハイレベル要件」[16] と「付属文書 3：認証ルールブック」[17] の 2 文書については特に技術仕様と規格の参考になるものとされている。

なお、この文書の位置づけは、eIDAS 専門家グループの進行中の作業の現状を示すものであり、その内容等に関して必ずしも正式に合意しているものではないこと、開発作業等を通じ時間の経過とともに補完され、更新されることには留意を要する。以下では、この EUDIW ARF1.4 版の内容をもとに EUDIW の仕様概要について読み解くことにする。

14：Commission Recommendation (EU) 2021/946 of 3 June 2021 on a common Union Toolbox for a coordinated approach towards a European Digital Identity Framework (Official Journal of the European Union, L 210, 14.6.2021, p. 51).

15：EUDIW ARF の目的は、eIDAS 規則を実施するために欧州委員会が策定する技術仕様、基準、手順であって、特に新たに改訂された以下のトピックに関連するものを定義することであるとしている。
EUDIW コア機能（第 5a 条）、EUDIW の依拠当事者（第 5b 条）、QEAA の要件（第 45d 条）、真正な情報源に対する属性の検証（第 45e 条）、公的機関（PSB）が発行する／公的機関のために発行される EAA に関する要件（第 45f 条）、国境を越えた ID 照合（第 11a 条）、EUDIW の認証（第 5c 条）、認証 EUDIW リストの公表（第 5d 条）、EUDIW のセキュリティ違反（第 5e 条）、PSB が発行する／ PSB に代わって発行される EAA の要件 - 通知（第 45f 条）

16：EUDIW エコシステムのエンティティに対する要求事項を規定する文章。

17：PID（個人識別データ）、m DL（モバイル運転免許証）の認証に関する具体的な要件を記載している規則集。

3．EUDIW ARF1.4版のポイント

3.1 ユースケース

EUDIW ARF1.4 版では、EUDIW の優先度の高いユースケースの青写真を示すことによって、潜在的な強化領域を浮き彫りにしつつ、サービス設計を助け、ユーザー体験とサービス効率を向上させるツールとして活用してもらうことを狙っている。そのため、ユースケースを知ることで、EUDIW が備えるべき機能や実際の活用イメージを想定することが可能となる。

（事例①）オンラインサービスにアクセスするための本人確認と認証

EUDIW は主に、公共および民間の様々なオンラインサービスにおいて、依拠当事者（Relying Party）がサービスを提供する利用者の身元を確実に確認できるよう、高い保証レベル（Levels of Assurance）での安全なユーザー識別と認証を容易にするよう設計されている。複数の本人確認方法を使用でき、ユーザーは、オンラインで個人識別データ（PID）を共有する際に懸念しうるプライバシーとセキュリティに特に留意している。このシナリオは、有効なウォレットインスタンスの取得から、オンラインサービスのための識別と認証のプロセスまでが含まれるなど、ユーザーの視点から EUDIW のライフサイクル全体をカバーしている。

（事例②）適格な電子署名

EUDIW では、ユーザーが適格な電子署名または印鑑を作成できなければならない。この目標は、ローカル QSCD、または QTSP によって管理されるリモート QSCD の一部として、EUDIW の認証と署名 / 印鑑機能を使用することで実現できる。

（事例③）携帯運転免許証

運転免許証は、EUDIW にとって重要なユースケースである。ユーザーはモバイル運転免許証（mDL）を取得、保存、表示しておき、必要の都度、例えば交通警察等に提示することができる。なお、EUDIW を使って mDL 属性の交換と開示を行うには、近接技術（例えば、NFC、Bluetooth）が用いられるが、人間またはその支配下にあるデバイス等に mDL 属性を提示する監視フローのシナリオと、無人の機械に対し mDL 属性を提示する非監視フローのシナリオが想定される。

（事例④）仮名（ペンネーム）によるアクセス

基本的な仮名のユースケースは、実名の提示を受けなくても、サービス提供者があらかじめ把握しているユーザーや、以前にやり取りしたことのあるユーザーであること等を認識できるようにするためのものである。ただし、この仮名は、必ずしもすべてのユースケースに適合するように設計されているわけではない。

（事例⑤）e ヘルス

健康関連データへ簡単にアクセスできることは、国内外の両方において極めて重要である。患者サマリー、e 処方箋などへのアクセスを可能にすることも推奨されている。

（事例⑥）学歴・専門資格の証明

EUDIW は、学歴や専門資格等の教育関係の属性電子証明書（教育デジタル・クレデンシャル）のリポジトリであり、関連するサービス提供者との間でそれらを交換するための手段となりうる。例えば、デジタル卒業証明書は、検証可能な信頼できる利用可能なフォーマットで、他の教育機関や訓練機関または将来の雇用主等に対し、国境を越えて提示される可能性がある。

（事例⑦）デジタル金融

EUDIW は、デジタルファイナンスや決済等の金融サービスにおける厳格な顧客認証要件へ準拠する強力なユーザー認証機能を提供することができる。

（事例⑧）デジタル・トラベル・クレデンシャル

ICAO（国際民間航空機関）が、パスポート情報をスマートフォンなどのデジタルデバイスに保持するために DTC（Digital Travel Credential）の規格化を進めているが、EUDIW に対しても DTC プロバイダーが対応するフォーマットで DTC を発行することができる。

3.2 設計原則

EUDIW の設計においては、４つの設計原則（ユーザー中心主義、プライバシー、セキュリティ、国境を越えた相互運用性を重視した要件への準拠が保証される。）が掲げられている。

ユーザー中心主義：EUDIW は、ユーザー中心主義を設計の基本方針としている。これは、ユーザーのニーズと経験をすべての設計決定に優先することを意味しており、ウォレットは直感的で使い易く、既存のユースケースにシームレスに統合されるべきとしている。ユーザーは、どのデータが誰と共有されているかについての透明性のある情報とともに、自分のデータとプライバシーを明確にコントロールできる。また、ウォレットは様々な技術的背景や能力を持つユーザーにも対応できるよう、アクセシブルになっている。

相互運用性：EUDIW は、EU 域内の国境を越えてシームレスに機能することを保証している。相互運用性は、標準化されたプロトコルを通じた安全なデータ交換を促進し、信頼できるエンティティがクレデンシャルを簡単に検証できるようにするため、ユーザーは自由に移動しながら、電子政府プラットフォームからプライベートなオンライン交流に至るまで、様々なサービスにウォレットを利用することができる。

デザインによるプライバシー：EUDIW アーキテクチャは、デザインによるプライバシーの原則を体現している。必要なものだけが収集されることを保証（データ最小化の原則）し、どのデータが誰と共有されるかをきめ細かくコントロールできる権限をユーザーに与え、データがどのように使用され、保護されるかがわかるようにシステムに組み込まれている（透明性の確保）。

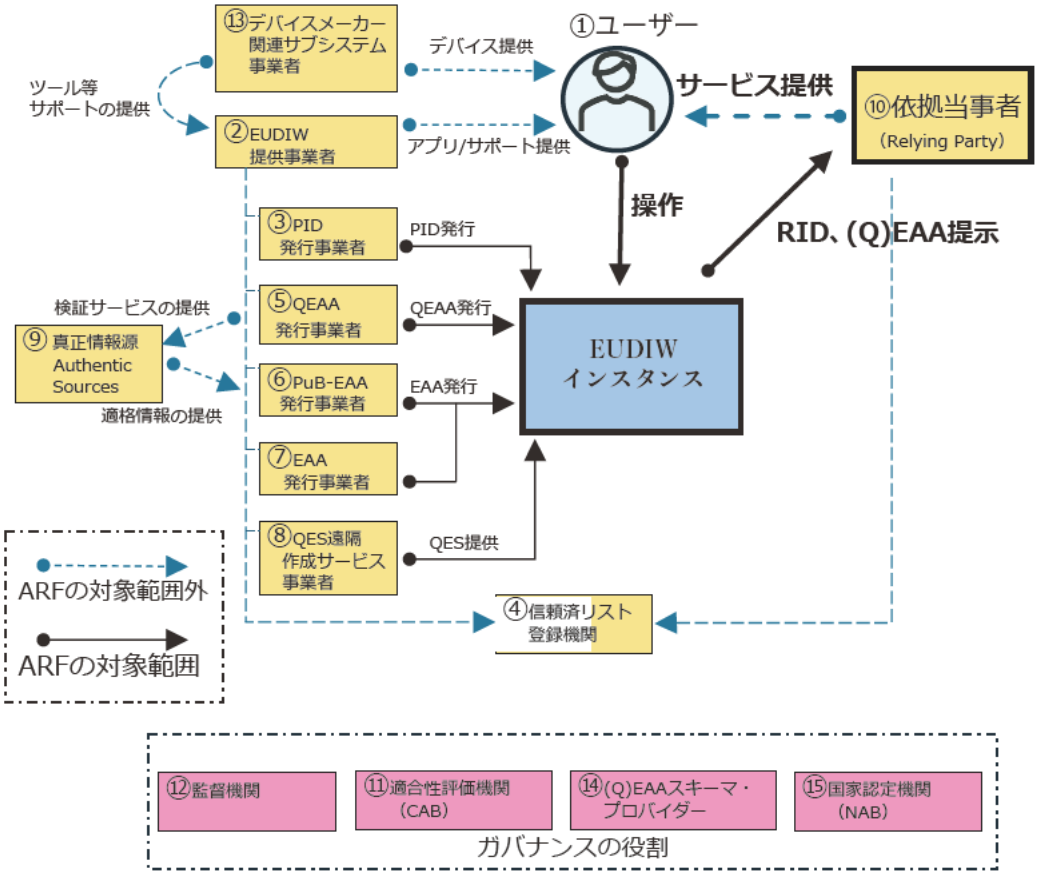
デザインによるセキュリティ：セキュリティへの配慮がウォレットの設計そのものに織り込まれている。設計プロセスを通して、潜在的な脆弱性への対応が行われているほか、セキュア・コーディングが義務付けられ、アーキテクチャ自体が機密データとアクセス制御を区分することで攻撃面を最小化している。

3.3 EUDIWのエコシステム

EUDIW ARF では、欧州デジタル ID 枠組み規則で想定されている EUDIW のエコシステムについて説明している。基本的な流れとしては、まず、ユーザーはデバイスメーカーやソフトウェア開発業者から EUDIW を入手し、その EUDIW のアプリ等を操作することで、個人識別データ（PID）や Attestation と呼ばれる用途に応じた様々な属性を認証できる電子形式の認証書（電子属性証明）をそれぞれの事業者から EUDIW に対して発行してもらっておく。そして、サービスを提供する依拠当事者（Relying Party）に対

して、EUDIW の中の必要な PID や電子属性証明のみを選択的に提示することで、必要なサービスを受けることになる。なお、こうしたエコシステムに存在するエンティティは、信頼済リスト登録機関に登録されていることが求められる。EUDIW のエコシステムにおける様々なエンティティとその役割の概要は図表 2 のとおりである。

図表2　EUDIWの役割の概要



出所）　筆者作成

① EUDIW のユーザー

EUDIW のユーザーは、EUDIW インスタンス[18] を使用して、PID や様々な電子属性証明を受信、保存、提示する。ユーザーはまた、EUDIW インスタンスにより適格電子署名 / 印鑑（QES）を作成し、ウォレット間のやり取りで利用することもできる。

② EUDIW 提供事業者（EUDI Wallet Provider）

EUDIW 提供事業者とは、EUDIW をエンドユーザーに提供する組織。その EUDIW ソリューションのインスタンスを通じてトラストサービス等を提供することで、利用者はその EUDIW 内の PID と電子属性証明、およびその他の個人情報の使用を完全に制御できるようになる（技術的には、関連する暗号秘密鍵等に対する制御権をユーザーが保有）。また、EUDIW 提供事業者は EUDIW が要件に準拠していることを保証する責任を持っている。

18：あらかじめ定義されたコンピュータプログラムやデータ構造などを、メインメモリ上に展開して処理・実行できる状態にしたもの。この文脈では「アプリ」が比較的近い意味。

③ PID 発行事業者（Person Identity Data Providers）

PID（個人識別データ）発行事業者は、以下の責任を負う信頼できるエンティティである。

・高い保証レベルの要求事項に従って EUDIW ユーザーのアイデンティティを確認する。
・EUDIW に統一された共通フォーマットで PID を発行する。
・依拠当事者（Relying Party）が PID の有効性を検証するための情報を提供する。

なお、PID 発行事業者は、例えば、公的身分証明書、電子身分証明書等を発行する組織や EUDIW 発行事業者と同じこともありうる。

④信頼済リスト登録機関（Trusted Lists Registrar）

信頼済リスト登録機関は、信頼済リストの維持、管理、公開に責任を持つ。EUDIW エコシステム内では、様々なエンティティに信頼済リストが存在する。信頼済リストには、主に関連するエンティティのトラストアンカー（ルート証明書）が含まれ、エンティティが作成した署名等の検証に用いられる。

⑤ QEAA 発行事業者（Qualified Electronic Attestation of Attributes Providers）

QEAA（適格電子属性証明）は、QTSP（Qualified Trust Service Providers：適格トラストサービス提供事業者）によって提供される。QEAA 発行事業者は、QEAA を要求・提供するために、EUDIW との相互認証インターフェースのほか、属性を検証するための真正情報源（Authentic Sources）とのインターフェースを保持する。なお、QEAA 発行事業者は、QEAA の有効性を照会するために必要な情報を提供するが、証明書の使用に関する情報は受け取らない。

⑥ PuB-EAA 発行事業者（Public Body Authentic Source Electronic Attestation of Attributes Providers）

PuB-EAA（公的電子属性証明）は、真正情報源に責任を負う公的機関等によって発行される電子属性証明である。 PuB-EAA の発行と運用に関する要件は、法的レベルで QEAA として認識できるようにするためのものである。

⑦（非認証）EAA 発行事業者（Non-Qualified EAA Providers）

EAA（非認証電子属性証明）は、どのような TSP（トラストサービスプロバイダー）からも提供される可能性がある。EAA は eIDAS 規則 のもとで監督されるが、EAA の提供、使用、承認に関する規則は、主に eIDAS 以外の他の法的または契約的枠組みによって規定されていることがほとんどであると考えられる（学歴証明書やデジタル決済など）。EAA を使用するためには、TSP はユーザーに対し、EAA を要求し取得する方法を提供しなければならず、必然的に EUDIW インターフェース仕様に準拠する必要がある。また、用途によっては、EAA 発行事業者は、EAA の有効性を照会するために必要な情報を提供することもできるが、ユーザー側の証明書の使用に関する情報は受け取らない仕組みとなっている。

⑧ QES 遠隔作成サービス事業者（Qualified Electronic Signatures Remote Creation Service Providers）

EUDIW を使用することで、ユーザーはあらゆるデータに対し無料で QES（適格電子署名）を作成することができる。これは、署名目的で EUDIW の普及が促されることに繋がる。

EUDIW を利用した電子署名の方法としては、「適格署名 / 印鑑作成デバイス」（QSCD: qualified signature/seal creation device）として認証されている EUDIW 自体で作成する方法のほか、QTSP（適格トラストサービス事業者）によって管理されるリモート QSCD 等の一部として、EUDIW に実装されている「セキュア認証と電子署名 / 印鑑の呼び出し機能」を用いる方法がある。図表 2 は、QES 遠隔作成サービス事業者がリモート QSCD になっている例である。

適格電子署名 / 印鑑の機能には共通のインターフェース／プロトコルが採用され、技術的相互運用性が確保されているため、リモート署名サービスを提供する QTSP の統一欧州市場が形成される。

⑨真正情報源（Authentic Sources）

真正情報源とは、法律で規定されている公的または私的なリポジトリないしシステムのことである。例えば、住所、年齢、性別、市民的地位、家族構成、国籍、教育・訓練の資格タイトルおよびライセンス、専門資格タイトルおよびライセンス、公的許可およびライセンス、財務および企業データ、などの情報源である。真正情報源は、QEAA プロバイダーに対し、上記属性の真正性を確認するためのインターフェースを、指定仲介機関を通じて提供することが求められるが、eIDAS 規則の要件を満たせば、自ら PuB-EEA を発行することもできる。

⑩依拠当事者（Relying Parties）

依拠当事者は、電子身分証明書またはトラストサービスに依拠することでユーザーを確認し、ユーザーに対して何らかのサービスを提供する自然人または法人である。 EUDIW の文脈では、依拠当事者は、ウォレット所有者であるユーザーの承認を条件として、適用される法律および規則の範囲内で、PID、QEAA、Pub-EAA、および EAA に含まれる必要な属性を要求する。EUDIW を活用したサービスの提供は、法的要件、契約上の合意、または依拠当事者自身の判断に基づくが、予め加盟国に対し設立場所とその利用意図を通知しておく必要がある。また、依拠当事者は相互認証で電子属性証明書等を要求するために EUDIW とのインターフェースを維持する必要がある。

⑪適合性評価機関（CAB：Conformity Assessment Bodies）

適合性評価機関（CAB）は、加盟国によって指定された国家認定機関によって認定された公的または民間の機関であり、EUDIW を発行する前、またはトラストサービスプロバイダーに認定ステータスを提供する前に必要な評価を実施する責任を負っている。例えば、EUDIW は CAB によって認証される必要があるほか、QTSP は CAB によって定期的に監査されることになっている。

⑫監督機関（Supervisory Bodies）

監督機関はウォレット提供事業者およびその他の関連団体の適切な機能を審査し、適切に機能していることを確認するために重要である。監督機関は加盟国に設置され、任命される。

⑬デバイスメーカーおよび関連サブシステム事業者

デバイスメーカーや関連サブシステム事業者などの商業主体は、EUDIW を安全かつ円滑に動作させるのに必要な、ハードウェア、オペレーティングシステム、セキュアな暗号機器、ライブラリ、アプリストア等のコンポーネントを提供する重要な役割を果たしている。

⑭（Q）EAA スキーマ・プロバイダー（Qualified and Non-Qualified Electronic Attestation of Attributes Schema Providers）

（Q）EAA の構造とセマンティクスを記述するスキーマと語彙を公開する主体が（Q）EAA スキーマ・プロバイダーである。これにより、依拠当事者など他のエンティティが（Q）EAA を検出して検証できるようになる。欧州委員会は、この目的のために最低限の技術仕様、標準、および手順を定めているが、セクター固有組織によるものも含む共通のスキーマは、（Q）EAA を広く普及させるために不可欠である。

⑮国家認定機関（NAB：National Accreditation Bodies）

国家認定機関は、加盟国由来の権限で認定を実行する加盟国の機関である。国家認定機関は、要件を規定する規範文書（法律、仕様、保護プロファイルなど）に照らして、製品 / サービス / プロセスを認証する責任を負う独立した監督下の専門認証機関として、適合性評価機関（CAB）を認定するとともに、これらを監視する。

EUDIW から依拠当事者に提示される情報には PID（個人識別情報）のほか、３つの電子属性証明書（Attestations）があるが、それらは以下のとおり区分され、法的に定義されている。

（EUDIWから提示されるAttestation等の種類）

―個人識別データ― PID（Person Identification Data）	欧州連合法または国内法に従って発行され、自然人もしくは法人、あるいは別の自然人もしくは法人を代表する自然人の身元を確認することができる一連のデータ。
―公的電子属性証明書― PuB-EAA（Electronic attestation of attributes issued by or on behalf of a public sector body responsible for an authentic source）	真正な情報源に責任を負う公的機関、または加盟国によって指定された代理の公的機関が発行する電子的な属性証明（Attestation）。
―適格電子属性証明― QEAA（Qualified Electronic Attestation of Attributes）	適格なトラストサービスプロバイダーによって発行され、規定される要件を満たす電子的な属性証明（Attestation）。
―非認定電子属性証明書― Non-Qualified EAA	QEAA でも PuB-EAA でもない EAA。

これらの認証の種類の違いは、純粋に法的なものであり、例えば、「卒業証明」は、適格なトラストサービスプロバイダー（QTSP）により発行されるか、非認定のトラストサービスプロバイダーにより発行されるかによって、QEAA となる場合もあれば、非認定 EAA となる場合もある。

3.4 EUDIWで扱うPID（個人識別データ）の概要

EUDIW ARF 文書では、PID、仮名、mDL、卒業証明書、電子処方箋などの各 attestation タイプについて、その証明の属性スキーマ、データ形式、証明メカニズム[19]、および認証と承認の信頼メカニズムを規定する認証ルールブックを定義することが要求されている。特に、組織間および / または国境を越えて使用されることを意図した認証ルールブックは、可能な限りすべての利害関係者が代表される組織によって定義されることになっており、これによって、同じタイプの認証（例えば、卒業証書）に対して複数の認証ルールブックが定義されることがなくなるとしている。

既に、PID ルールブック[20]、mDL ルールブック等に関するものは定義されているが、このルールブックによると、EUDIW で管理され依拠当事者へ提示できる個人識別データは次のとおりである。

19：属性スキーマは、認証された属性の構造、論理構成、タイプ、名前空間、および認証、発行者、検証メカニズム、基礎となる身元保証、プロパティが関連するトラストフレームワーク、正当なユーザーによる所有の証明などの追加情報である。データ形式は、文字セット、エンコード、シリアル化など、証明書のデータの形式を指す。証明メカニズムは、選択的開示も含め、完全性と真正性の証明に使用される方法を意味する。

20：ANNEX 3.1 - PID Rulebook https://github.com/eu-digital-identity-wallet/eudi-doc-architecture-and-reference-framework/blob/main/docs/annexes/annex-3/annex-3.01-pid-rulebook.md

図表3　PIDの概要

Attribute identifier（属性識別子）	定義	
family_name（姓名）	現在の姓または名	必須
given_name（名前）	ミドルネームを含む現在のファーストネーム	必須
birth_date（生年月日）	生まれた日、月、年	必須
age_over_18（18歳以上）	現在成人（true）か未成年（false）	必須
age_over_NN（年齢）	NN 歳以上か?	任意
age_in_years（年齢）	現在の年齢	任意
age_birth_year（年齢_誕生年）	生まれた年	任意
family_name_birth（姓名_出生）	出生時の姓または名	任意
given_name_birth（出生）	出生時の姓名（ミドルネームを含む）	任意
birth_place（出生地）	生まれた国、州、都市	任意
birth_country（出生国）	生まれた国。ISO 3166-1 alpha-2国コード	任意
birth_state（出生地）	生まれた州、県、地区、または地域	任意
birth_city（出生地）	生まれた市町村	任意
resident_address（居住者住所）	現在居住している、または連絡が取れる場所の完全な住所（通り名、家屋番号、市町村など）	任意
resident_country（居住国）	現在居住している国。ISO 3166-1 alpha-2国コード	任意
resident_state（居住州）	現在居住している州、県、地区、または地域	任意
resident_city（居住都市）	現在居住している市町村	任意
resident_postal_code（居住者郵便番号）	現在居住している場所の郵便番号	任意
resident_street（居住ストリート）	現在居住している通りの名前	任意
resident_house_number（住居番号）	現在居住している家の番号	任意
Gender（性別）	性別。ISO/IEC 5218の定義値	任意
Nationality（国籍）	国籍、ISO 3166-1 alpha-2国コード	任意
issuance_date（発行日）	PIDが発行された日付	必須
expiry_date（有効期限）	PIDの有効期限が切れる日付	必須
issuing_authority（発行機関）	このPIDインスタンスを発行した行政当局の名前等	必須
document_number（ドキュメント番号）	PIDプロバイダーによって割り当てられたPIDの番号	任意
administrative_number（管理番号）	監査管理などの目的でPIDプロバイダーが割り当てる番号	任意
issuing_country（発行国）	PIDプロバイダーの国または地域。ISO 3166-1 alpha-2国コード	必須
issuing_jurisdiction（発行管轄）	PIDを発行した法域の国細分コード。ISO 3166-2:2020の第8節で定義	任意

出所）　筆者作成

これを見ると、生年月日に関連するものだけでも、以下の属性が定義されている。

必須	birth_date（生年月日） age_over_18（18 歳以上か?）
任意	age_birth_year（誕生年） age_in_years（年齢） age_over_NN（NN 歳以上か?）

生年月日について 1 つの属性だけでなく粒度の異なる複数の属性を持つことで、ニーズに応じてリクエストやレスポンスで使い分けることができ、PID プロバイダーと依拠当事者は保有データを必要最小限に抑えることができるとしている。例えば、いくつかのユースケースにおいて、依拠当事者は PID ユーザーが未成年者でないことを証明するだけでよく、その場合、age_over_18 を要求すれば十分である。PID ユーザーの年齢や誕生年など、より具体的な情報を公開することは、ユーザーのプライバシーを不必要に侵害することになる。

本文書では age_over_18 を必須属性とし、その他の age_over_NN 属性（NN は年齢）を任意属性としているが、PID プロバイダーは複数の age_over_NN 属性を自由に追加することができる。

このほか、出生地関連属性としても、「国、州、都市」、「ISO 3166-1 alpha-2 国コード」、「州、県、地区、または地域」、「市町村」と異なる粒度のものが存在するほか、アドレス（住所）関連属性についても完全な住所を含む 7 種類の属性が定義されるなど、複数の属性を持つようになっている。

なお、PID は相互運用性を高める観点から標準フォーマットが決められており、ISO/IEC 18013-5:2021[ISO18013-5] で規定されている形式と、[SD-JWT VC] で規定されている形式の両方で発行されなければならないと規定されている。前者の場合、属性は CBOR でエンコードされなければならないほか、後者の場合、属性は JSON でエンコードする必要がある。

4．おわりに

わが国のマイナンバーカードは、対面で公的な身分証明書として使う以外に、IC チップ部分に保存された電子証明書を用いて、オンラインでの本人認証にも使うことが可能である。マイナンバーカードでの健康保険証利用も始まり 2024 年 12 月から一体化が決まっている。自動車運転免許証についても、2022 年 4 月に公布された改正道路交通法でマイナンバーカードに免許関連の情報を電磁的に記録する規定が整備され、2024 年中にはマイナンバーカードの運転免許証としての利用がスタートするといわれている。

一方、マイナンバーのスマートフォン対応については、Android スマホでは 2023 年 5 月から、マイナンバーカードの「電子証明書」の機能を内蔵する「スマホ用電子証明書搭載サービス」が開始されている。さらに、「マイナンバーカードのすべての機能をスマートフォンに搭載できるようにする」マイナンバー法の改正案が 2024 年 5 月 29 日、国会で成立した。一部の認証サービス提供企業では、「スマホ用電子証明書搭載サービス」を自社サービスとして既に提供し始めている。iPhone では対応が遅れていたが、これにより「スマホ用電子証明書搭載サービス」のみならずマイナンバーカードの「券面記載事項」（氏名、生年月日、住所、性別、マイナンバー、顔写真等）の搭載も可能となり、来年春には iPhone の Apple ウォレットにマイナンバーカードを追加することでアップル社と合意している。機能的には、本人が所有する国家資

格証明書などもスマホ画面上で提示できるようになるとの見方もある。

これらの動きを見ると、結果的に、わが国でも EUDIW と同様の取組みが行われているようにも見えるが、マイナンバーカードの機能拡張の文脈で検討されているものであり、必ずしもデジタル ID ウォレットのあるべき姿をデザインして導かれた姿ではない。そのため、

- 十分考えられたエコシステムが存在しない中で官民、特に民間での活用に制限がある、
- 技術仕様や共通ルールの透明性が不十分であり、相互運用性に懸念がある（欧州のような国境を跨る運用も想定されていない）、
- 現在搭載されている電子証明書は「署名用電子証明書」と「利用者照明用電子証明書」の 2 種類のみであり、プライバシーに配慮して制御可能な細かい属性認証（例えば「18 歳以上」等）への対応の検討が遅れている、

といった課題がある。

こうした課題を含めマイナンバーカードの抱える問題に対しては、これまでのマイナンバーカードにはグランドデザイン（全体設計）が欠如していたことが問題であったと指摘する専門家[21]も散見される。これに対しては、先行して進んでいる EUDIW の取組みをもとに、日本版デジタル ID ウォレットのグランドデザインを検討することが解決に繋がる可能性があると思われる[22]。そして、マイナンバーカードがデジタル ID ウォレットとして進化を遂げ、官民で使えるオンラインサービスが拡充されるなど高い利便性が伴ってくれば、わが国特有のマイナンバーに対する誤解や不信も薄まるのではないだろうか。

なお、EUDIW ARF では、EUDIW エコシステムのコアとなるコンポーネントを概説した「リファレンス・アーキテクチャ」や、それらのコンポーネント間で安全なやり取りを可能とするために、当事者間でどのような信頼関係が確立されるかを記述した「トラストモデル」についても、多くのページが割かれているが、大部である他かなり技術的な内容になるため本稿では割愛した。これについては、機会があれば改めて解説したい。

21：一般社団法人 情報システム学会 マイナンバー制度研究会が公表した「「マイナンバー制度の問題点と解決策」に関する提言の補足」（2024/7/3）の付属文書「やさしい解説：マイナンバー制度のあるべき姿とは」でも、「問題の根本は、目的に到達するためのグランドデザインすなわち全体像を描いた制度設計がないまま、マイナンバーカードの普及に力点を置いてデジタル化を推進してきた点にあります。政府はマイナンバーカードを「デジタル社会のパスポート」と呼んで、次から次へと機能を追加しています。一方、デジタル先進国で日本のような多機能 IC カードを発行している国は見当たりません。むしろ IC カードを使わないでデジタル化を推進し成果を上げた国の方が多いくらいです。」としている。

また、国際政治学者の舛添要一氏も、「マイナンバーカードですべての用事が済むようにするためには、カードを構想する段階から、グランドデザインが必要である。例えば、納税データとの連結などがそうである。しかし、このようなデザイン設計を最も苦手とするのが役人である。官庁の縦割り、縄張り根性も邪魔になる。最初から民間の優秀な専門家に任せていれば、こうはならなかったであろう。」と主張している。

22：デジタル庁が公表した「次期個人番号カードタスクフォース最終とりまとめ」（2024/3/18）によると、2026 年を視野に導入の検討が進められている次期マイナンバーカードでは、諸外国 eID カードの事例を参考に一部の機能を見直そうとしていることが読み取れるが、エコシステムを含むグランドデザインの検討に関する記述はみられない。

参考文献

European Commission（2021), COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS 2030 Digital Compass: the European way for the Digital Decade COM/2021/118 final, 2021/3/9, https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0118

――――――――― (2021), Commission Recommendation (EU) 2021/946 of 3 June 2021 on a common Union Toolbox for a coordinated approach towards a European Digital Identity Framework, Official Journal of the European Union, L 210, 2021/6/14, p. 51

――――――――― (2023), The European Digital Identity Wallet Architecture and Reference Framework v1.0, 2023/2/10, https://digital-strategy.ec.europa.eu/en/library/european-digital-identity-wallet-architecture-and-reference-framework

――――――――― (2024), European Digital Identity Wallet Architecture and Reference Framework v1.4.0, 2024/5/23, https://eu-digital-identity-wallet.github.io/eudi-doc-architecture-and-reference-framework/1.4.0/arf/#314-qualified-and-non-qualified-electronic-attestation-of-attributes-schema-providers

――――――――― (2024), European Digital Identity Wallet Architecture and Reference Framework v1.4.0 ANNEX 3.1 - PID Rulebook, 2024/5/23, https://github.com/eu-digital-identity-wallet/eudi-doc-architecture-and-reference-framework/blob/main/docs/annexes/annex-3/annex-3.01-pid-rulebook.md

European Parliament, Council of European Union (2022), Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030, Official Journal of the European Union, L323, 2022/12/19, p. 4

――――――――――――――――――――――――――― ,European Commission (2023), European Declaration on Digital Rights and Principles for the Digital Decade, Official Journal of the European Union, C23, 2023/1/23, p. 1

――――――――――――――――――――――――――― (2024), Document 32024R1183, Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing the European Digital Identity Framework, 2024/4/11, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1183

一般社団法人情報システム学会マイナンバー制度研究会（2024）「『マイナンバー制度の問題点と解決策』に関する提言の補足―やさしい解説：マイナンバー制度のあるべき姿とは」、2024 年 7 月 3 日、https://www.issj.net/teigen/2024_myno_kaisetu.pdf, p. 1

デジタル庁（2024）「次期個人番号カードタスクフォース最終とりまとめ」、2024 年 3 月 18 日、https://www.digital.go.jp/assets/contents/node/basic_page/field_ref_resources/58b82d5b-338d-4f5b-be7e-7b771135e2c3/0d7a6b39/20240318_meeting_mynumber-card-renewal_outline_06.pdf

舛添要一（2023）「混迷するマイナカード、『グランドデザイン』描けない役人に最大の責任がある」Shirabee ニュース、2023 年 6 月 18 日、https://sirabee.com/2023/06/18/20163098937/#

伝統的金融に呑まれる分散型金融

― 暗号資産 ETFと合同会社型 DAOを例に考える ―

斉藤 賢爾｜早稲田大学大学院経営管理研究科教授

要約

2024 年 1 月から 5 月にかけて、米国にて BTC や ETH といった暗号資産の現物価格を指標とする ETF（Exchange Traded Funds）が承認されたり、我が国において合同会社型 DAO（Decentralized Autonomous Organization）が法的に認められたりといった規制緩和が行われ、一般公衆が暗号資産に基づく分散型金融（Decentralized Finance）にアクセスしやすくなったとして注目を浴びている。これらの新しい金融商品や組織形態は、伝統的金融と分散型金融の双方の歩み寄りの結果のようにも見えるが、伝統的金融との接点を必ず必要としていた暗号資産交換業と同じ問題を抱えることになるのではないだろうか。本稿では、暗号資産とブロックチェーンのそもそもの成り立ちを振り返りながら、これらの新しく見える概念について、その意味を改めて議論する。そのことを通して、最初から伝統的金融なくしては成り立たず、伝統的金融に呑まれていた分散型金融の姿が見えてくる。

斉藤 賢爾

早稲田大学大学院経営管理研究科教授

コーネル大学より計算機科学において工学修士号、慶應義塾大学よりデジタル通貨の研究で博士号を取得。日立ソフトウェアエンジニアリング、慶應義塾大学大学院政策・メディア研究科特任講師等を経て現職。

1．はじめに

2024 年 1 月と 5 月、米国では証券取引委員会（SEC）が Bitcoin（Nakamoto, 2008）のネイティブ暗号資産[1]である BTC と Ethereum（Buterin, 2013）のネイティブ暗号資産である ETH の現物価格を指標とする ETF（Exchange Traded Funds; 上場投資信託）を相次いで承認した[2]。一方、同年 4 月、我が国では金融商品取引法に関わる内閣府令の改正により、法人格を持てる合同会社型 DAO（Decentralized Autonomous Organization; 分散型自律組織）が可能となった。

これらによって、一般の消費者が適切な保護や法的な裏付けの下で暗号資産に基づく分散型金融（Decentralized Finance）や分散型自律組織にアクセスできるようになると見込まれることから、こうした最近の動きは多くの人々により好意的に迎えられているように見える。この現象は、伝統的金融が分散型金融に歩み寄ったということなのだろうか、それとも分散型金融が伝統的金融に呑み込まれたのだろうか、あるいはその両面があるのだろうか。

1：本稿では、ブロックチェーンの維持活動に参加することで新規発行分を得られる暗号資産を、そのブロックチェーンのネイティブ暗号資産と呼ぶ。各ブロックチェーンではそうではないデジタル資産を定義し、その所有権の移転を記録することもできる。

2：ETH の現物 ETF に関しては、2 段階の承認プロセスの最初の段階が 2024 年 5 月に承認されたのであって、最終的な上場承認はこの稿の執筆時点では未だ成されていなかったが、同年 7 月に承認された。

本稿では、暗号資産とブロックチェーンのそもそもの成り立ちから振り返ることを通して、暗号資産の現物価格を指標とする ETF や合同会社型 DAO といった新しく見える概念について、その意味を改めて議論する。これらの金融商品や組織形態は、伝統的金融と分散型金融の双方の歩み寄りの結果のように見えるかも知れないが、その背景には、元々伝統的金融なくしては成り立たないという、分散型金融の根幹が見落とされていると筆者は考える。本稿は、最終的に「分散」と「集中」の対立軸に焦点を当て、分散型金融が本来持つ意味とその実状を露わにする。

2. 暗号資産の成り立ちを振り返る

2.1 設計のゴール

サトシ・ナカモトを名乗る人物ないし集団が、最初のブロックチェーンであり暗号資産システムである Bitcoin について初めて記述した設計文書（Nakamoto, 2008）は、その冒頭で、信用できる第三者としての金融機関を通した送金[3] を問題視していた。そのような第三者は原理的に送金を検閲でき、口座の凍結といった形で資金の移動を否定できるからである。したがって、Bitcoin の設計のゴールは、そのような第三者を排した、何人によっても検閲できない送金システムであり、「自分が持つ資金を自分が選んだ誰かに自由に送るのを誰にも止めさせない」ことの実現だと考えられる。

この要求は、次の 4 つの性質に分解できる。

1. 自己主権性 —利用者自身の意思のみによって利用者はシステムに参加でき[4]、送金を指示できる。

2. 狭義の耐検閲性 —利用者が指示する送金は、他の誰の意思によっても止められない。

3. 耐障害性 —利用者が指示する送金は、システムの故障・障害によっても止められない。

4. 耐改ざん性 —送金の記録は後から削除・変更できないし、過去に無かった送金の記録も捏造できない。

これらすべてを満たすことで、広義の耐検閲性（何人・何事によっても記録を否定できない）が実現される。

2.2 設計の概要

どのような技術であれ、無条件にその力を発揮できるわけではなく、例えば腕時計には、正常に動作する温度範囲が（あくまで例として）-10 ～ 60℃といった仕様がある。ブロックチェーンも同様に、特定の条件下で上の 4 つの

3：オンラインシステムでは支払いも送金によって実現されるので、支払いと送金は区別されない。

4：誰の許可も得ずとも自らの意思だけでアカウントを作れることを意味し、典型的には公開鍵と秘密鍵の鍵ペアを自分（のウォレット）で生成し、公開鍵をアドレスに変換し、秘密鍵を用いた演算（デジタル署名）によりそのアドレスが示す本人であることの認証を行う。

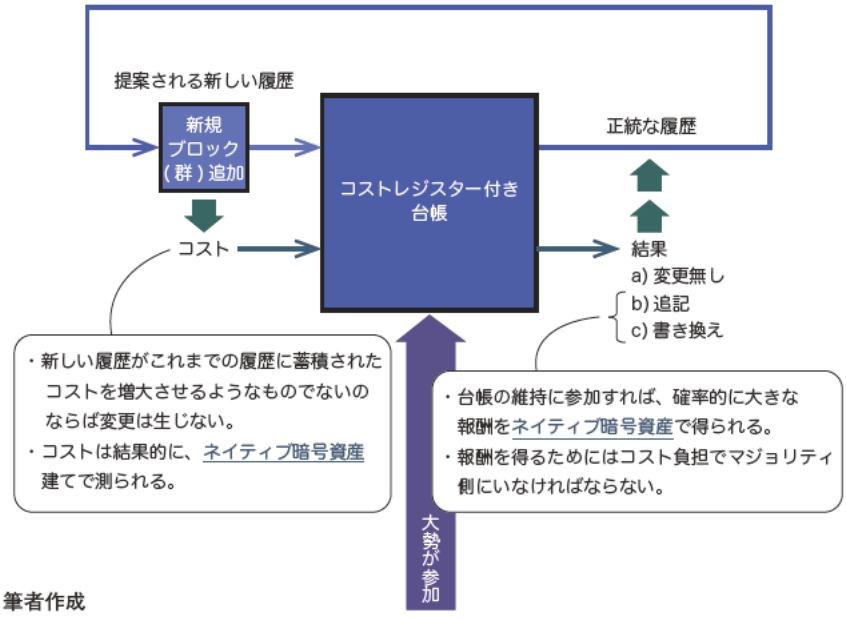
性質を満たすように設計された技術であって、条件から外れた状況では広義の耐検閲性が満たされるとは期待できない。では、ブロックチェーンの動作条件とは何だろうか。

そのことに留意しながらブロックチェーンの仕組みを説明するために、筆者が（Saito and Yamada, 2016）にて導入した、「コストレジスター付き台帳」というメタファーを用いることにしたい（図 1）。これは、台帳（ブロックチェーン）への記録が行われる際に投入されたコストをレジスター（登録器）に加算していき、蓄積されたコストが現状よりも大きくなる、すなわち記帳には追加コストが必要であり、それを誰かが実際に負担した場合のみ新たな記録を許す仕組みで、正統な送金履歴を維持する回路のようなものである。この台帳ないし回路は、大勢がその維持のために参加するインセンティブ構造を持つ仕組みとなっている。

利用者たちによる送金の記録はそうした大勢によって集められ、ブロックと呼ばれる構造にまとめられ、台帳に書き込まれる。ブロックは大勢のうちのひとりだけが作成し提案できる。送金履歴はブロックの列として表現されるが、これがブロックチェーンという名前の由来である。

新たなブロックが提案され追加される際には、それを認める大勢により大きなコストが負担される。 Bitcoin に代表される Proof of Work 方式の場合、そのコストはブロックの提案者を選ぶくじ引きのために各々が負担する電力コストであるが、後述するように各参加者は報酬による収益を期待してそのコストを負担するため、その上限はネイティブ暗号資産建て報酬の市場価格の期待値となる。 Ethereum に代表される Proof of Stake 方式の場合、そのコストは各々がネイティブ暗号資産建てで負担するデポジットおよび各自が正統と見なす履歴に対する投票であり、正統な履歴はデポジット換算で全体の$\frac{2}{3}$以上の票を集め続ける必要がある。

図1　暗号資産の成り立ち



出所）　筆者作成

コストレジスター付き台帳は、ブロックが追加された新しい履歴とそれに伴うコストを入力として受け取り、結果（変更無しか、履歴に追記するか、あるいは履歴を書き換えるか）と新たな正統な履歴を出力する。新しい履歴のコストが、現在の正統な履歴のコストを上回らなければ（すなわち、履歴更新にかかるコストが正でなければ）、変更は生じない。変更は、典型的にはひとつのブロックの追記だが、ある特定の過去から最新までの一連のブロックが改めて提案し直され、それによって履歴が置き換わる（reorg、すなわちチェーンの再構成と呼ばれる）場合もありうる。

満たすべき ４つの性質のうち、「自己主権性」はアドレスの設計により担保され、「狭義の耐検閲性」と「耐障害性」は、大勢が独立した運用方針をもって各々のコンピューターを参加させ、その結果、誰かが止めたい取引でも他の誰かによって記録が行われたり、誰かのコンピューターが止まっていても他のコンピューターが動いていたりといった冗長性がもたらされることによって担保される。そして「耐改ざん性」は、改ざんしようとする主体が単独では履歴の書き換えのためのコストを負担できないことにより担保されるため、コスト負担ベースで悪意はマイノリティなことが前提となる（下線部がブロックチェーンの技術的な動作条件であり、これらが成立するための動機的な条件を後述する ）。

台帳の維持に参加すると、確率的に大きな報酬を、そのために無から生み出されたネイティブ暗号資産で得られる。報酬を得るためにはコスト負担でマジョリティ側にいる必要があるが、このことでマジョリティが形成されやすくなり、履歴がまとまりやすくなる。

以上では台帳には送金が記録される前提で説明したが、台帳にはプログラムコードやその呼び出し、その実行結果を記録することもできる。それがスマートコントラクトである。スマートコントラクトを金融に応用したものが分散型金融だと言える[5]が、本稿では加えてネイティブ暗号資産の取り扱いも分散型金融の概念に含めている。

5：ブロックチェーンの ４つの性質を満たさない、プライベートな台帳上でのコード実行も含めて分散型金融だと主張する向きもあるため、この定義は狭いものだと言える。

2.3 暗号資産と DAO

Ethereum の共同作者のひとりであるヴィタリック・ブテリンは、ブログ記事（Buterin, 2014）にて DAO をこう定義した。「インターネット上に自律的に存在するが、自動システム自身にはできない特定のタスクを担うために人間を雇うことに大きく依存しており、そのために内部に資本（報酬として使われ人間を駆動する）を持つ。」

これは、ある意味ブロックチェーンのイメージそのものであり、ブロックチェーンこそが DAO の原型であるとすら言える。ブロックチェーンは、無からネイティブ暗号資産を生み出し、それを資本として参加者である大勢の人間を金銭インセンティブを通して雇い、駆動し、かつネイティブ暗号資産建てで計量されるコストによって台帳を改ざんから守っているからである。

このことが成立するためには、ネイティブ暗号資産が市場で十分に高値を付けている必要がある。安ければ大勢の人間を駆動できないし、ブロックチェーンを改ざんするためのハードルが低くなるからである。ブロックチェーンが正

常に動作するための動機的な条件は「ネイティブ暗号資産の市場価格が十分に高いこと」なのである。

2.4 価格形成

それでは、ネイティブ暗号資産の市場価格はどのように形成されるのだろうか。商品の価格は、一般に需要と供給のバランスによって決まる。需要が高まれば価格が上昇し供給が増えて需要とマッチし、需要が低まれば価格が下降し供給が減ってやはり需要とマッチする。需要の変化に対しては、一般に供給が反応することで価格の大きな変動は抑えられる。

一方、ネイティブ暗号資産に関しては、こうした供給の調整が成り立たないことを筆者らは Iwamura et al.（2019）および Saito and Iwamura（2019）にて議論した。ネイティブ暗号資産の新規供給はブロック提案の報酬として行われ、ブロックは一定の時間間隔で提案される設計であるため、需要の変化に対して供給は反応しない。したがって、需要が高まるほど価格は上昇し、需要が低まるほど価格は下降してしまう。

加えて、Bitcoin では 21 万ブロック毎に報酬が半減するいわゆる半減期を通して新規供給を減らし、最終的には全体の供給量が固定になることを目指している。また、筆者らが Saito et al.（2023）にて議論したように、Ethereum には半減期は無く、参加者数に応じて一定の割合で ETH が新規供給されるが、トランザクション（取引）の実行の基本手数料として支払われる ETH がバーン（消滅）されることを通して全体の供給量を一定に保とうとしている。需要が高まり、多くのトランザクションが発生すると、逆に全体の供給量は減ることになる。

さらに問題になりうるのは、暗号資産は産業と無関係に売買される点である。といっても、Bitcoin に代表される Proof of Work 方式の場合、くじ引きの計算の効率化に向けて半導体事業への投資を促し、また、より効率的な発電方法や、批判を避けるべく環境負荷の低い発電方法への移行を促すとも言える。対して Ethereum に代表される Proof of Stake 方式はネイティブ暗号資産によるデポジットによってシステムを保護する考え方であり、産業との接点を持たない。スマートコントラクトも主として分散型金融分野のアプリケーションであり、その多くも同様に産業との接点を持たないとすれば、単に需要を高めれば手持ち資産の価格が上がり、需要を低めれば手持ち資産の価格が下がる装置として利用されるに過ぎないことになってしまう。

3. 伝統的社会からの歩み寄り

3.1 前提となる接点

ブロックチェーンが正常に動作するための動機的な条件は「ネイティブ暗号資産の市場価格が十分に高いこと」なのだから、ネイティブ暗号資産の取引市場が存在する必要がある[6]。

そのため、多くの暗号資産交換業者が国内外で事業を行うニーズがあるわけ

6：2009 年に Bitcoin が稼働を始めた時点ではそのような市場は無かったが、技術的な興味によってブロックチェーンが維持されていたと考えられる。

だが、多くはドルや円といった旧来の通貨で暗号資産の売買が行われており、暗号資産の価値評価において、米ドルといった伝統的金融における通貨建ての価値が参照されている。この意味で、暗号資産の取引市場は分散型金融と伝統的金融との接点だと言える。

こうした多くの事業では、取引所・販売所の顧客間や交換業者との間で暗号資産の残高を付け替えることで取引が実行され、顧客が明示的に自分のウォレットへの引き出しを指示するまで、実際にはブロックチェーンには暗号資産の所有の移転は書き込まれない（ブロックチェーン上の取引手数料を節約し、かつ高速に売買を成立させるため）。ということは、こうした市場で売買されているのは、暗号資産の現物により裏付けられた何か、すなわち、交換業者によって管理される、ブロックチェーンとは別物の台帳の上に記録された暗号資産の所有残高である。そう考えると、すでに現物 ETF に相当するような抽象化が日常的に行われていたとも言える。

また、ブロックチェーン自体をその原型としていたと考えられる DAO も、「自動システムによって人間が雇われる」という部分が「雇われて業務を執行している社員たち自身により組織が所有されている」といったように解釈が緩和・拡張され、スマートコントラクトによる経営ルールと、同じくスマートコントラクトにより生成されたガバナンストークン[7] の持ち分に応じた投票権により組織の意思決定を行うタイプの DAO が乱立するに至った。これは基本的にはトークンの二次市場での値上がりを期待する勢力が参加者の大部分を占め、暗号資産の取引市場と同じマインドセットにより DAO が支配されてしまう類のものである。こうした現象は、元々の DAO の概念が伝統的社会における既存の株式会社や合同会社の概念に引っ張られて変化したものだと考えられる。

こうした動きを支えているのが、いわゆる「クリプト界隈」と呼ばれる、暗号資産の利用者の集合でありコミュニティである。クリプト界隈と一般の投資家たちの間には、リスクの捉え方にギャップがあり、暗号資産自体のボラティリティの大きさもさることながら、取引所・販売所のセキュリティーのリスク（サイバー攻撃や、詐欺や価格操作など）への寛容度の違いが、一般投資家の参加へのハードルとなっている。

7：トークンは代替貨幣の意味で、ここではスマートコントラクトにより生成されたデジタル資産を表す。

3.2 暗号資産の現物価格を指標とする ETF

2024 年 1 月 10 日、米国証券取引委員会（SEC）は BTC の現物 ETF を承認した（SEC, 2024a）。ETF の発行者は暗号資産取引市場から BTC を購入し、カストディ（管理・保管）サービスを通して保全し、ETF の裏付け資産とする。市場での BTC の購入と売却を通して、BTC 現物 ETF の放出と回収は BTC の市場価格に影響を及ぼしうる。

SEC は、同様の ETF の提案を、詐欺や価格操作のリスクを理由に長らく不承認としてきた。しかし、コロンビア特別区（ワシントン D.C.）巡回控訴裁判所が、そうした提案のひとつである Grayscale Bitcoin Trust（GBTC）に対する SEC の決定を無効とする判決を 2023 年 8 月 29 日に下したため、再検討を余儀なくされた（Gensler, 2024）。その再検討において、SEC は、

CME（Chicago Mercantile Exchange）の BTC 先物市場といった、現物市場と十分に相関関係が見られる取引所との監視共有契約がある場合は、提案された BTC の現物 ETF に関連する潜在的な詐欺や操作を監視する適切な手段が確保されていると判断した。これらの提案が市場の透明性、公平性、および効率性を促進すると同時に、投資家を詐欺や価格操作のリスクから保護することができるという信念のもと、承認の決定が行われたわけである。

同年 5 月 23 日、SEC は ETH の現物を裏付け資産とする ETF を承認した（SEC, 2024b）（上場に至る承認プロセスの第 1 段階）。

この場合においても、CME の ETH 先物市場と現物市場間の相関に基づく詳細な相関分析が提供され、SEC は、ETH の現物 ETF 市場での不正や操作が CME の ETH 先物価格に同様の影響を与える可能性が高いと結論付け、その結果、CME との包括的な監視共有契約が、提案された ETH 現物 ETF に影響を及ぼす不正や操作を監視する上で有効に機能すると期待したのである。

このように、暗号資産の現物 ETF は、暗号資産の現物に裏付けられた何かを売買するという暗号資産取引市場の考え方をある意味踏襲したまま、それをさらに一般の投資家が参加する伝統的金融の世界に引き込むべく、安全性に関わる手当てをしたものだと言える。

3.3 合同会社型 DAO

LLC（Limited Liability Company; 有限責任会社または合同会社）を DAO 化し、既存の法体系にあまり手を加えずに DAO を法人化する考え方は、米国ワイオミング州（Wyoming, 2022）をはじめとして各地で見られる。ワイオミング州の法律では、既存の LLC を DAO に転換することもでき、LLC が DAO と互換性があると考えられていることを示している。

我が国においても、2024 年 4 月 22 日に施行された「金融商品取引法第二条に規定する定義に関する内閣府令の一部を改正する内閣府令」および改正された「金融商品取引法等に関する留意事項について（金融商品取引法等ガイドライン）」（金融庁，2024）に拠り、合同会社型 DAO の設立が可能となった。

合同会社は、所有と経営が分離されていない会社（持分会社）の形態のひとつであり、出資者（社員）が経営を行い、原則として全員一致で定款変更などを行い、社員自らが会社の業務を執行する。合同会社型 DAO は、既存の合同会社の枠組みにブロックチェーン技術を組み合わせ、社員権を表すトークンやその他のトークン（別トークン）を発行することで資金調達ができ、また、トークンを保有する社員やその他の参加者による、例えば投票を用いたガバナンスを可能としている。

一般に DAO はガバナンストークンの買い占めによるテイクオーバー（買収）が可能であるが、合同会社型 DAO では業務執行社員権トークンの譲渡は相手も業務執行社員である場合にのみ可能であるので、合同会社と同様（というよりも法的にも合同会社であるので）、歓迎しない買収のリスクを避けられる。また、社員全員が自分の出資分だけの責任を負う有限責任を持つし、中でも債務者から直接弁済の追及を受けない間接有限責任を負うことで守られる。

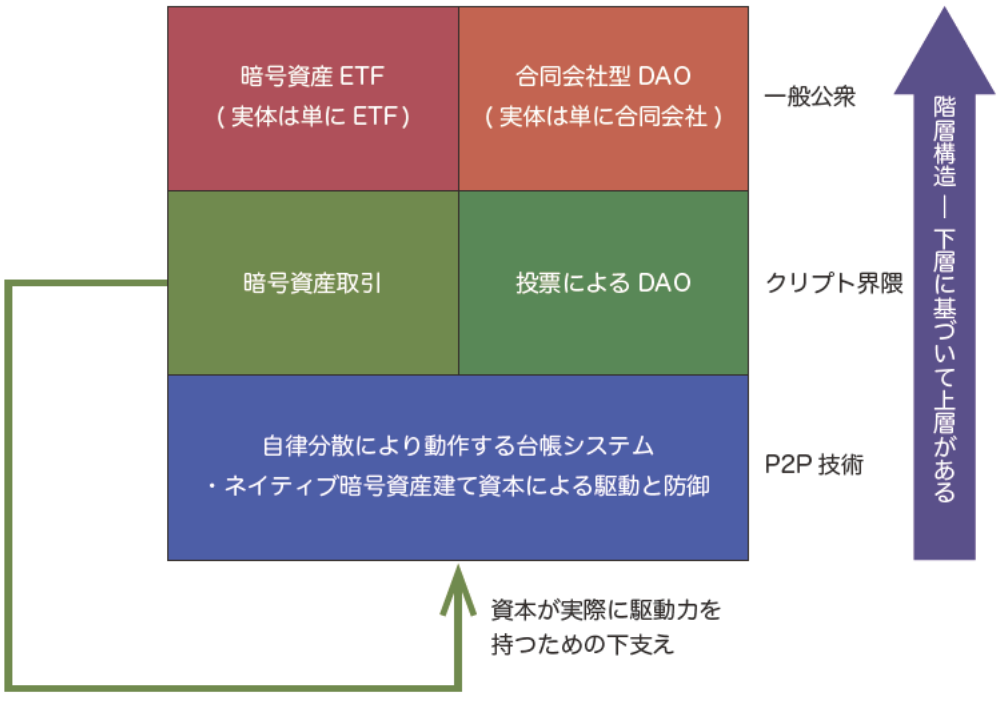
このように、合同会社型 DAO は、ブロックチェーンのインセンティブの構造として登場した原型としての DAO を株式会社・合同会社寄りに拡大解釈した「投票による DAO」を、さらに既存の合同会社として法的に解釈可能にし、法による保護を手当てしたものだと言える。

4. 分散型金融は伝統的金融に呑まれるのか

4.1 最初から呑まれていた分散型金融

さて、以上のように、暗号資産の現物 ETF や合同会社型 DAO といった新しい概念は、一般の消費者・投資家に対して十分な保護を提供することで、暗号資産へのアクセスを容易かつ安全にした。このことの背景には、図 2 に示すような階層構造がある。階層は、下層に基づいて上層があることを示す。

図2　暗号資産と DAOの階層構造



出所）　筆者作成

P2P（Peer-to-Peer）技術、すなわちネットワークに参加するコンピューターが対等に役割を担い、故障や妨害に強い柔軟なサービスを構築できる仕組みの賜として、ブロックチェーンという、自律分散により動作する台帳システムが作られた。そして、ブロックチェーンの上でネイティブ暗号資産が作られたり、スマートコントラクトを実行できることが前提となって、暗号資産取引や投票による DAO が可能となった。また、暗号資産取引市場が存在することが前提となってその ETF 化が可能となったし、投票による DAO が合同会社との類似性をもつことから合同会社型 DAO、すなわち社員権のトークン化が可能となった。これらの実体はそれぞれ単に伝統的社会における ETF であ

るし、合同会社（の社員権）である。

ここで忘れてはならないのは、ブロックチェーンはネイティブ暗号資産建て資本によって人間を雇い、駆動し、またネイティブ暗号資産建てで計量できるコストの高さによって守られているのだから、暗号資産取引市場（によってネイティブ暗号資産に高値が付けられること）に依存しているということである。

先に述べたように、暗号資産取引は伝統的金融との接点であり、ネイティブ暗号資産に付けられる高値は伝統的金融の価値で測られるのであるから、分散型金融は伝統的金融に最初からすでに呑まれていたのである。

4.2 考えうるリスク

原理が異なるものに呑み込まれているのだから、分散型金融の挙動は、ブロックチェーンが正常に動作するための条件から外れる恐れがあるし、伝統的金融の期待にも応えられない局面が生じうる。

その例としてリスク分散がある。ブロックチェーンのネイティブ暗号資産は需要が高まれば価格が上がり、低まれば価格が下がるのだから、単に別種の金融商品に投資をする選択肢として暗号資産を選ぶだけではリスクが分散されるとは限らない。例えば、株式市場と暗号資産取引市場の値動きに負の相関があるという期待があるとする。株の値段が下がるときに、株を売って暗号資産を買うという行動が需要の趨勢を決定づけるのであれば、実際にその局面で暗号資産の価格が上がる。これは予言の自己成就のメカニズムであり、最初に予言があり、その予言が市場の参加者に周知されていることが必要なのである。

SBI 金融経済研究所が行ったアンケート調査の結果（SBI 金融経済研究所, 2022, 2024）では、暗号資産に関する認識を問う設問に「投資対象を値動きの異なる金融商品に分散することで、投資のリスクを低減する効果がある」という選択肢があり、暗号資産への投資に関するリスク分散の可能性についての認知が一部の回答者にはあるものの、全体的な浸透度は必ずしも高くはないことが示された。このことは、仮に今後、暗号資産の現物 ETF が広く投資家の関心を集めたとして、人々が思うほどにはリスク分散の効果が出ない可能性にも留意すべきことを示唆しているのかも知れない。

また、現物 ETF の承認が各ブロックチェーンのネイティブ暗号資産価格の上昇に寄与するという期待は理解できるが、逆に将来的にそうした暗号資産の価格が暴落する可能性も無視できない。以下はそのようなシナリオを引き起こす可能性のある要因である。

- 市場の過熱：ETF の承認が大きな期待を生み、短期間で暗号資産の価格が急騰する可能性がある。しかし、このような急騰はしばしば市場の過熱を示唆し、商品貨幣のように本源的価値が存在せず、発行体に対する請求権がある負債性資産でもない暗号資産に市場価格が付くこと自体がバブルであり、過去の暗号資産価格の暴騰暴落を見ても、投機的バブルの発生の可能性は高い。期待が現実に見合わない場合、その後の価格調整は急激な暴落を引き起こす可能性がある。

- 流動性の問題：ETF への投資が増えると、その裏付けとなる資産への需要も増加する。しかし、ある時点で、市場における売り手が買い手を上回り、ETF 市場とその対象資産市場（暗号資産市場）の両方において流動性の問題が生じる可能性がある。これは価格が急落する要因となりうる。
- 規制リスク：金融商品としての ETF の承認により、暗号資産への規制当局の注目が高まる可能性がある。将来的に、より厳しい規制が導入されると、市場の不確実性が増し、投資家の信頼が損なわれ、結果として価格が下落する可能性がある。
- 技術的な課題：将来の暗号技術や量子アルゴリズムの飛躍的な進歩や、現時点で未知の脆弱性を突いたサイバー攻撃等により、ブロックチェーンのセキュリティーに関する未解決の問題が露呈する可能性がある。これらの技術的な問題が表面化すると、投資家の信頼が揺らぎ、価格が下落する可能性がある。
- 市場心理：投資市場での心理的な要因は価格に大きな影響を及ぼす。一旦、市場センチメントが転換し、恐怖が支配すると、パニック売りが起こり、価格の急落を引き起こすことがある。
- マクロ経済要因：グローバル経済の状況、金利の変動、通貨の動向など、暗号資産市場の外の要因も価格に大きな影響を及ぼす。これらの要因が不利な方向に進むと、暗号資産全体の価格下落を引き起こす可能性がある。

ただし現状では、価格が下降すれば、将来の上昇を見込んで買う投資家が現れる。このことで、これまでは主要な暗号資産については存続が危ぶまれるほどの過度な暴落は避けられてきた経緯がある。

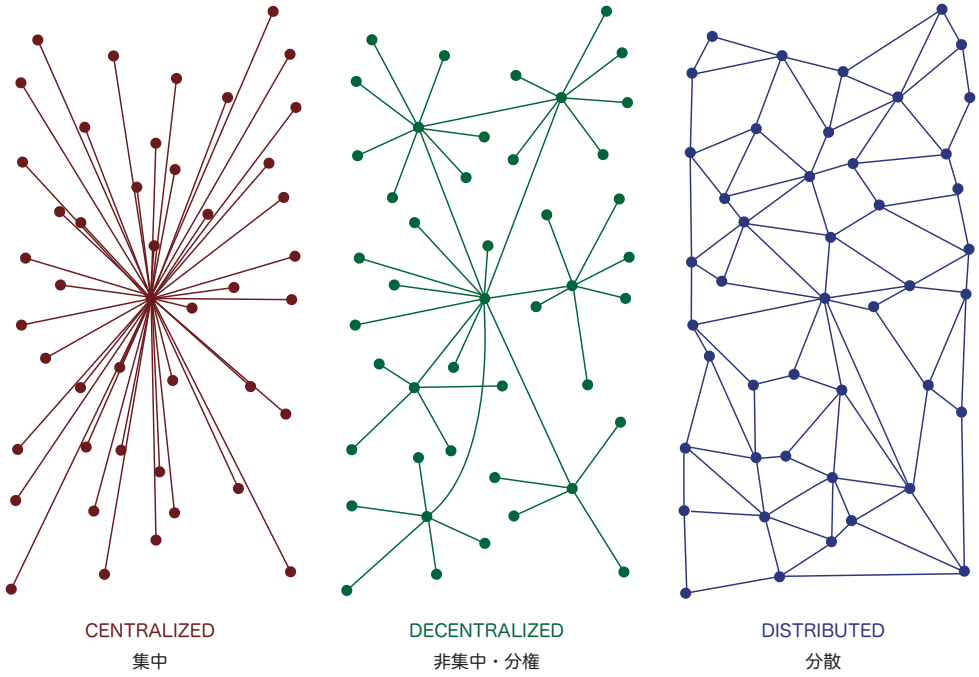
しかし、ブロックチェーンの原理に照らせば、ネイティブ暗号資産の市場価格の低下はそれ自体がリスクなのだから、実際にブロックチェーンの維持参加者の撤退が起きてもおかしくはない。暗号資産取引市場とブロックチェーンには、原理の乖離がある。

実際にブロックチェーンが維持できなくなった時、下層である基盤が揺らぐのだから、暗号資産の市場自体が崩壊すると考えられる。

おわりに

いわゆるクリプト界隈では、DAO や DApps（Decentralized Applications）といったように、“Decentralized”という言葉が好んで用いられる。これは、おそらくは“Centralized”（集中・中央集権）な伝統的金融に対するアンチテーゼとしての言葉の用い方で、Decentralized Finance に対する訳語「分散型金融」が示しているように、意図としては「分散」という概念を表したいのだろう。しかし、Baran（1964）に見られるように、コンピューターサイエンスの分野では、遅くとも 1960 年代から、“Decentralized”は多くの中心に権限が分かれる「非集中・分権」を表す言葉として明確に定義され、用いられてきた（図 3）。

図3　集中（centralized）、非集中･分権（decentralized）、分散（distributed）



出所）　Baran（1964）　Fig.1に基づいて筆者作成

P2P 技術の賜である台帳システムのネットワークは、特定の構造を持たずに実際に分散（distributed）（図 3 の右）なのだが、暗号資産取引や投票による DAO のネットワークは、期せずして様々な取引所・販売所やスマートコントラクトといった多数の中心を持つ構造であり、非集中・分権（decentralized）（図 3 の中央）である。それが悪いわけではないが、考えていることと実体の間に乖離がある。そしてこの構造は、多数の金融機関のネットワークが相互に接続された伝統的金融と同じ形なのである。

分散を志しながらも、伝統的金融との接点を通してしか自らを維持できない台帳システム自体に問題の根幹はあると言えるのかも知れない。何人・何事によっても記録を否定できないような台帳の存在自体は有益であるので、そろそろ、伝統的金融には依存しない、別の動作条件の下で動作する台帳システムに向けて、私たちの開発の労力を振り向ける時機が到来したと言えるのではあるまいか。

謝辞

この稿の執筆に当たり議論とインスピレーションをもたらした、早稲田大学大学院経営管理研究科「ブロックチェーンと分散ファイナンス」ゼミ（2024年度）の学生諸氏にこの場を借りてお礼を申し上げたい。

参考文献

Baran, P. (1964). On Distributed Communications: I. Introduction to Distributed Communica.tions Networks. RAND Corporation, Santa Monica, CA.

Buterin, V. (2013). A Next-Generation Smart Contract and Decentralized Application Platform. https://ethereum.org/en/whitepaper/.

———— (2014). DAOs, DACs, DAs and More: An Incomplete Terminology Guide. https://blog.ethereum.org/2014/05/06/daos-dacs-das-and-more-an-incomplete-terminology-guide/.

Gensler, G. (2024). Statement on the Approval of Spot Bitcoin Exchange-Traded Products. https://www.sec.gov/newsroom/speeches-statements/gensler-statement-spot-bitcoin-011023.

Iwamura, M., Kitamura, Y., Matsumoto, T., and Saito, K. (2019). Can We Stabilize the Price of a Cryptocurrency?: Understanding the Design of Bitcoin and Its Potential to Compete with Central Bank Money. Hitotsubashi Journal of Economics, 60(1).

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. http://bitcoin.org/bitcoin.pdf.

Saito, K. and Iwamura, M. (2019). How to make a digital currency on a blockchain stable. Future Generation Computer Systems, 100:58.69.

Saito, K., Soejima, Y., Sugiura, T., Kitamura, Y., and Iwamura, M. (2023). Is Ethereum Proof of Stake Sustainable? － Considering from the Perspective of Competition Among Smart Contract Platforms －. https://arxiv.org/abs/2309.11394.

Saito, K. and Yamada, H. (2016). What's So Different about Blockchain? Blockchain is a Probabilistic State Machine. In 2016 IEEE 36th International Conference on Distributed Computing Systems Workshops (ICDCSW), pages 168.175.

U.S. Securities and Exchange Commission (2024a). Order Granting Accelerated Approval of Proposed Rule Changes, as Modified by Amendments Thereto, to List and Trade Bitcoin-Based Commodity-Based Trust Shares and Trust Units. https://www.sec.gov/files/rules/sro/nysearca/2024/34-99306.pdf.

———— (2024b). Order Granting Accelerated Approval of Proposed Rule Changes, as Modified by Amendments Thereto, to List and Trade Shares of Ether-Based Exchange-Traded Products. https://www.sec.gov/files/rules/sro/nysearca/2024/34-100224.pdf.

Wyoming Secretary of State, Business Division (2022). Decentralized Autonomous Organization (DAO): Frequently Asked Questions. https://sos.wyo.gov/Business/Docs/DAOs FAQs.pdf.

SBI 金融経済研究所（2022）.「次世代金融に関する一般消費者の関心や利用度に関するアンケート調査」結果 . https://sbiferi.co.jp/questionnaire/question20221227.html.

————（2024）.「次世代金融に関する一般消費者の関心や利用度に関するアンケート調査、第 2 回」の結果 . https://sbiferi.co.jp/questionnaire/question20240425.html.

金融庁（2024）.「金融商品取引法第二条に規定する定義に関する内閣府令の一部を改正する内閣府令（案）」等に対するパブリックコメントの結果等について . https://www.fsa.go.jp/news/r5/shouken/20240401/20240401.html.

生成AIウォークスルー：基本技術、LLM、アプリケーション実装

副島 豊｜SBI 金融経済研究所研究主幹　兼　SBI ホールディングス
SBI 生成 AI 室プロジェクトコーディネーター

要約

本稿は、生成 AI とりわけ大規模言語モデル（LLM）の発展と実装技術に関する展望論文である。ニューラル言語モデルの基礎、LLM に繋がる理論モデルの発展、様々な LLM の群雄割拠、発展過程で発見されてきた多様な転移学習の形態、それがもたらした利用法の拡大、企業での活用法（秘匿情報を扱う手法）、実装に必要な技術群とその学び方、最後に、実際の構築事例を紹介している。生成 AI 技術の活用には手を動かして学ぶことが重要であり、6 つの学習ステップをその一例として示している。LLM の仕組みやシステム実装に関する本稿の情報は、生成 AI 技術を金融実務に活用していく際の学びや実践の水先案内（Pilot）となろう。

副島 豊

SBI 金融経済研究所研究主幹　兼
SBI ホールディングス SBI 生成 AI 室プロジェクトコーディネーター

1966 年生まれ。京都大学卒、90 年日本銀行入行。フィンテックセンター長や金融研究所長を歴任。90 年代より様々な先進的分析手法を日本銀行に導入。金融システムレポートや各種レポートを企画・創刊。BIS・グローバル中央銀行活動のエキスパートメンバーとして国際基準策定等に参画。

1．はじめに

ChatGPT が彗星のように登場し、人々の耳目を集めたのは 2022 年末であった。僅か 1 年半前のことである。これまで自然言語処理（NLP: Natural Language Processing）や AI 技術とは縁がなかった人々が、文章でコンピュータに指示を出し、その都度コンピュータによって生成された文章で回答を得るという新しい体験を享受した。コンピュータが文章や画像、音声などの情報を直接処理し、オリジナルなものを新たに生成するという新技術の普及は、ビジネスや日常生活、社会活動に無限の可能性をもたらす。多くの企業や個人、公的セクターが生成 AI 技術の有用性を強く意識するようになり、生成 AI は 2023 年に最も注目された言葉の一つとなった。金融ビジネスにおいても、対顧客サービスや企業内利用において様々な活用法が検討され、実際にサービスとして提供され始めている。

一方で、生成 AI を検索サービスのように利用する誤用法や、これに伴うハルシネーション（幻覚、もっともらしい嘘）が話題となった。また、生成 AI の学習時における利用情報や、そのプライバシー保護、情報セキュリティ管理、生成内容の公正性や倫理面での問題[1] など、様々な課題も意識されるようになった。

1：宇根（2022）は、金融サービスの提供において機械学習による予測・推論を利用する際の公平性・公正性に関する諸問題を俯瞰している。同研究の対象は機械学習であるが、生成 AI にとっても同様な問題が存在している。

生成 AI サービスの創造と活用の促進、および上述のような課題への対応においては、①生成 AI がどのような技術であるか、その中身を知り、②どのような IT インフラの上に構築されているか、構築していけばよいのかを知り、③急速に進化を遂げる生成 AI 技術のフロンティアをキャッチアップし続けること、が必要となる。①の理解が不足していると、生成 AI の利用が適切でないタスクを選択したり、機械学習など他の手法が適切な分野に生成 AI を応用してしまい、望ましい成果を得られない危険性がある。②の理解が不足していると、アイディアはあっても実用化に至らないか、陳腐化が著しい（寿命が短い）生成 AI 活用サービスにおいて外注を繰り返すことになる。ビジネスモデルの DX（デジタルトランスフォーメーション）を巡る議論では、デジタル社会におけるサービス構築や業務構築には IT 知識が必須であり、内製化やビジネス推進の現場が IT を学ぶ必要性がより強く意識されるようになってきている。生成 AI も同様であり、現場が技術を理解していないと、優れたサービスを発想し実装することは難しい。また、③の対応をとらないと、より高度な生成 AI 技術やそのサービス実装方法が次々に誕生し、機能・コスト・実装の容易さが劇的な速度で改善していくもとで、競争優位性があるサービスを提供することが困難となる。本稿は、金融ビジネスに携わる企業やビジネスパースンが、上記①～③にかかわる基礎的な知識を習得できるよう、生成 AI 技術の中核となっている LLM（大規模言語モデル）の原理と発展、実装方法をできるだけ平易に俯瞰することを試みている。

学術研究においても生成 AI の活用が生産性や競争力を高めるうえで重要となっている。知的創造活動は生成 AI の恩恵を最も享受しやすい分野である。その高い情報集積・処理能力は、論文執筆における文献レビュー、研究対話や議論などの壁打ち、仮説生成、データ分析、分析プログラム作成、論文構成、文書校正などに有益である。例えば、文献レビューを効率化する生成 AI アプリケーションサービスとして Google 提供の notebookLM がある[2]。同サービスは、RAG と呼ばれる生成 AI の活用法（本論文で解説）を極めて簡便に利用できるようにした生成 AI アプリケーションサービスである。PDF や Web ページで提供される情報をまとめてデータベース化し、この情報に基づいた質疑や節ごとの要約、理解度チェックなどが利用できる。本年 6 月より日本語対応版が提供され始めており、Google 開発の LLM 最新モデル Gemini 1.5 Pro がバックエンドで稼働している。6 月末現在で Experimental 版として無償提供されている。

図表 1 は、筆者の最近の論文、副島（2024）を PDF アップロードして情報データベース化した画面である。論文の概要が示されており、目次タブをクリックすると各節の要約が LLM によって作成され、簡便に文献レビューを行うことができる。下方には、論文に関する質問を受け付ける窓がある。回答は論文情報の範囲内で作成されるため、ハルシネーションが生じにくくなっている。作成根拠となった論文中の箇所が示されるため、回答内容の確認や論文中の当該箇所へのアクセスも行いやすい。学習ガイドのタブでは用語集や、内容理解度を確認できるクイズ（解答付き）などが提供される。その一部を図表 2 に示した。

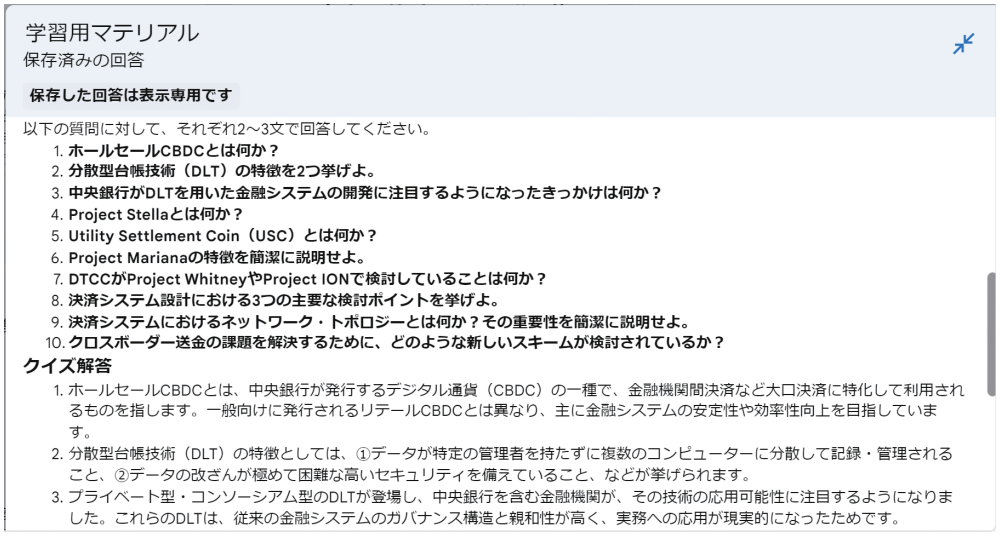
2：ChatGPT は、その機能をカスタム化して特定のサービス構築を可能とするプラットホーム GPTs を提供している。同サービス上では、Consensus や Scholar GPT、Paper interpreter など文献検索やレビューに特化した便利なサービスが提供されている。ChatGPT の有償ユーザになると各種の GPTs を使用できる。GPTs サービスの作成者は自作サービスが大量利用された場合は、これを収益化することができる。

図表３は、SBI 金融経済研究所の Web サイトで直近１年間に公表された Web レポート 24 本を html アドレス指定で読み込ませて情報データベースを作成した画面である。ある研究分野の代表的な論文を大量にデータベース化すると、同分野に精通した生成 AI サービスを作成することができる。こうした有益なサービスが次々に誕生しており、生成 AI の開発動向や用途の発展を知ることは研究者にとっても重要となっている。

図表１　notebookLMの画面：副島（2024）論文の読み込み

出所）筆者作成

図表２　学習ガイドで自動作成されたマテリアルの一部：理解度クイズ



出所）筆者作成

図表3　SBI金融経済研究所Webレポートの情報データベース化

出所）筆者作成

本稿の構成は以下のとおりである。２節では LLM の基礎となったディープニューラルネットワーク（DNN）型の自然言語モデルについて解説する。3節では、自然言語モデルの発展について解説する。特に、Transformer モデルの登場によって急速な性能向上がもたらされ、その後に LLM 時代が到来したことや、これが現在の最新 LLM にどう展開していったかについて解説する。４節では、LLM 発展の過程でスケール則や創発現象、転移学習が発見され、これらが LLM 開発の方向や応用手法を決めていったことを紹介する。5節では、非公開情報を利用する生成 AI サービスの主要な開発手法となっている RAG（Retrieval Augmented Generation）について、追加学習を行う Fine-tuning と対比させながら解説を行う。６節では、生成 AI アプリケーションを実装する方法について、IT インフラや実装技術の進化の面から解説する。専門家でないと携わることができなかったアプリケーションのシステム実装において、参入障壁を引き下げる民主化の波が押し寄せていることを簡単な RAG の構築事例を示しながら紹介する。最後に７節では、LLM の学際的研究の可能性として、人間の認知や行動の理解という視点から、World Models や人工市場シミュレーション、経済理論モデルへのインプリケーションを議論する。

2．生成AIの中核技術LLMを支える基礎技術

自然言語処理の研究は長い歴史を持つ。言語をコンピュータで解析し理解する技術（自然言語理解、NLU）として、言語の文法や規則性を研究しルール化することでテキストを解析する手法などが 1950 年代から考案されてきた。文章を単語の固まりに分解する形態素解析や、固有名詞などの固有表現抽出、文章のツリー構造を判別する構文解析、述語項構造認識[3]、コーパス[4] や類義語辞書などの作成などである。テキスト分析やテキストマイニングも NLU の一分野であり、コーパスなどから自動的に情報を抽出・分析し、知識を整理・発見したり、仮説を検証する研究などが行われてきた。文書の分類（トピック分類や性質分類 < 例えばスパム / 非スパムメール >）、ポジティブ・ネガティブ等の感情分析、書き手の判別（紫式部の文章か否か）などである。言語の生成アルゴリズムを対象とした自然言語生成（NLG）の研究により、機械翻訳や文書要約、対話の創造などの手法が発展してきた。コンピュータが一般化した 1980 年代以降は NLP に統計的手法が導入され、大規模なコーパスを用いた単語の登場頻度や順位、単語間の関連性（隣接して出現する共起発生の頻度 <n-gram> など）の研究が行われた。

ニューラルネットを応用したニューラル言語モデルの登場（Bengio et al. 2003）によって自然言語処理へのアプローチが変わり始めた。そして、2017 年に登場した Transformer モデルにより、言語理解や生成のパフォーマンス（例えば機械翻訳の精度）の面で、過去の長い研究蓄積を短期間のうちに凌駕していった。本節では、まず、ニューラル言語モデルが確率モデルとして文章を生成していく方法を解説し、次に、ニューラル言語モデルの基礎となっているニューラルネットワークについて説明する。最後に、言語をコンピュータで処理できるよう数値化する手法、特に膨大なコーパスを活用して単語間の関係性を抽出した数値化を行う手法（単語の分散表現 / 埋め込み表現）を解説する。

2.1 ニューラル言語モデルの確率的文章生成

ある単語の並び、例えば、{日本 / の / 首都 / は /} があったとする。次にくる単語を予想すると、「東京」が最有力候補であろう。もちろん、「暑い」でもよいし、「京都 / だった」が続くこともありうる。しかし、蓋然性が最も高いのは東京であろう。確率言語モデルも同様な発想に基づく。条件付き確率 P（東京 | {日本 / の / 首都 / は /}）を計算し、東京以外の他の全ての単語候補と比較し、一番確率が高いものを選択して、これまでの文章にその一単語を付け足す。これを次々に繰り返すことで文章が生成されていく。

広辞苑の収録数は約 7 万語である。固有名詞を含めると日本語にはそれより遥かに多くの単語が存在する。例えば、日本語 Wikipedia をコーパスとして単語数を計測した事例を見ると 130 ～ 150 万の範囲で報告されている（web クロールのコーパスを含めると 200 万）。これらのすべてについて上記のような確率計算を行うと膨大な計算負荷が生じる。このため、様々な計算節約手法が考案されているが、文章生成の原理は上述の通りであり、最も高い条件付き

3：誰がいつどこで何をしたという術語に関わる構造を文章から取り出す（認識する）こと。

4：自然言語として書かれた文章や使い方を大規模に収集し、コンピュータで検索できるよう整理されたものをコーパスと呼ぶ。文章を構造化したうえで品詞など言語的な情報を注釈として付与したものが一般的であり、文章をそのまま集めたものは生コーパス（raw corpus）と呼ばれる。文書の多様性を考慮して構築された均衡コーパス（balanced corpus）、複数言語の翻訳をセットにした対訳コーパスなど、様々なコーパスが作成されている。かつては紙で作成されていたが、現代ではコンピュータで扱えるようデジタル情報化されている。これにより統計処理、例えば、ある文書における特定単語の登場頻度に注目した TF-IDF 分析などが容易に行えるようになった。詳細は岡崎他（2022）を参照。

確率を示した単語を選択するか、多少のゆらぎをいれるかなどのバリエーションが選択・調整可能な仕組みが取り入れられている[5]。

こうした言語生成の仕組みは、その用途ほどには知られていない。ChatGPT の登場により LLM に社会的な注目が集まったが、当時、LLM を検索サービスのように誤用することが広範に生じた。LLM は学習用の膨大なコーパスに存在している単語間の登場のパターン性を学習し、次の単語を確率的に選択し繋ぎあわせることで文章を生成しているにすぎない。にもかかわらず、整理された正しい知識や情報が LLM から引き出せるかのように利用され、結果、ハルシネーションを起こす点が問題視された。確率モデルとしてのニューラル言語モデルの成り立ちを理解すると、ハルシネーションの発生は必至であることがわかる。

その後、LLM が高度化・巨大化し、学習のためのデータセットも膨大なものとなったため、あたかも LLM が膨大な知識を保有しているかのように見える度合いが格段に高まった。一方で、単語間の関連性や登場順、組み合わせなどに関するパターン性が様々なトピックについて極めて精緻に把握され、その結果、正確な情報を提供する能力が高まってきており、LLM に知識が蓄えられたと捉えることもできる。このため、上述のような誤解が解消されにくくなっており、これが LLM の用途の適切な選択に悪影響を及ぼしている面がある。

しかし、ここ 1 ～ 2 年のモデル規模や学習情報の巨大化により、知識データベースとしての有用性が加速的に高まっている。例えば、MMLU（Massive Multitask Language Understanding）という言語モデルの汎用性、理解力、推論能力を評価するための包括的なベンチマークテストがある。多言語かつ様々な領域の知識理解を一般常識から専門知識まで幅広く試すものである[6]。その評価スコアによると、最先端の LLM は一般人を遥かに上回る専門家並みの知識を有しているという指摘がなされている[7]。同テストは選択形式であり、回答文書生成時のハルシネーション発生リスクは別途存在しているが、言語生成における汎用能力も改善を続けており、公開情報の範囲内での知識に基づく言語生成に限定すればハルシネーションのリスクは低下しつつある。また、ハルシネーションを回避する技術的な工夫もなされている。例えば、RAG を活用して根拠となるオリジナル情報の該当箇所を提示したり、プロンプト技術（入力する指示情報に詳細な条件を付加することで意図した結果を引き出す工夫）によってハルシネーションを予防する工夫などである。このほか、特定分野の専門情報を追加学習させることによって知識を強化することで一般情報に基づく誤答の可能性を回避するような LLM も開発されているため、知識データベース的な活用が更に進むと予想される。本稿の執筆においても、最新の LLM 群や LLM 活用サービスを情報検索などに活用している。

2.2 ニューラルネットワークの仕組み

LLM の基盤となっているモデルはニューラルネットワーク（以下、NN）である。1980 年代前半に登場し、第二次 AI ブームを牽引した NN は、脳のシナプスの構造を単純化して表現したモデルである。図表 4 のようにインプッ

5：例えば、①確率が一番高いものを逐一選ぶ（greedy decoding）、②複数個先までのセットで確率を評価する（beam search）、③確率的ゆらぎをいれる（random sampling）といった方法がある。③では、上位 p%（あるいは k 個）の候補からランダムに選択する方法や、temperature という手法が知られている（確率のばらつきの差を人為的に縮小させることでゆらぎが起こりやすくするもので、GPT が temperature や p% 法を採用している）。①は分類問題、②は機械翻訳、③は長文生成でしばしば利用されている。

6：各種の LLM リーダーボードでも評価基準として活用されている。3 節では日本語性能を評価している LLM リーダーボードを紹介している。

7：Hendrycks et al.(2021)、OpenAI et al.(2023) を参照。

トとアウトプットの間を中間層によって繋ぎ、アウトプットが正解（教師データ）と合致するようパラメータを調整する仕組みである[8]。下層の要素を線形和し、これを非線形変換したものが上層の要素となる。これはシナプスが他の多数のシナプスからの電位差を受け取り、その合計値がある閾値に近づくと急速に（非線形的に）出力が上昇して発火する（新たな電位差を作り、隣接するシナプスに伝える）仕組みを模倣したものである。

線形和をとる際の係数や定数項（バイアス）、非線形変換のパラメータを変えることで、図表４中の４つのインプットは中間層の各要素（ノード）に異なる値をもたらす。中間層からアウトプットに向けても同じ仕組みが適用される。インプット変数の組み合わせ比率を変えることで中間層の各ノードに異なった情報を集約し、インプットとアウトプットの間にある何らかの関係性をノードが分担しながら捉えようとするものである。これは、機械学習における特徴量抽出と類似した機能である。こうしたNNは、推論時にはインプットからアウトプットに向かって前向きに情報が伝わっていくため、フィードフォワード型NNと呼ばれる。NNの最も基本的なモデルとなっている。

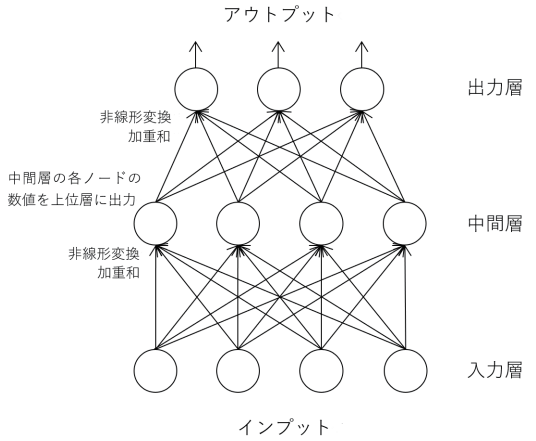
非線形関数には、図表５で示したようなシグモイド関数、ハイパーボリック・タンジェント関数、ReLU（Rectified Linear Unit）やその変形判が用いられる[9]。シグモイド関数の場合、アウトプットが１つであれば０か１という判別問題に対応でき、選択肢が３つであれば、アウトプットを３つ（y_1, y_2, y_3）とし、教師データを（1,0,0）、（0,1,0）、（0,0,1）とすることで判別問題に対応できる。なお、1要素だけが1という値をとり、他要素がゼロであるベクトルを、1要素だけが発火しているという意味でone-hotベクトルと呼び、ニューラル言語モデルでも重要な役割を担っている（2.3節の分散表現で後述）。

このようにインプットとアウトプットを繋ぐのは四則演算やシンプルな非線形関数に過ぎない。しかし、各要素は単純な関数であっても、インプット数や中間層の要素数を増やし、中間層を複数積み重ねる（図表４では１層のみ）ことで、複雑に機能する非線形関数を作りだすことができる（今泉 2021）。

8：この推計をNNでは学習と呼ぶ。学習には、アウトプットと教師データの誤差最小化問題が利用され、微分による漸近的なパラメータ調整を、アウトプット層から中間層、インプット層に向かって進めるため、誤差逆伝播法（バックプロパゲーション）と呼ばれる。

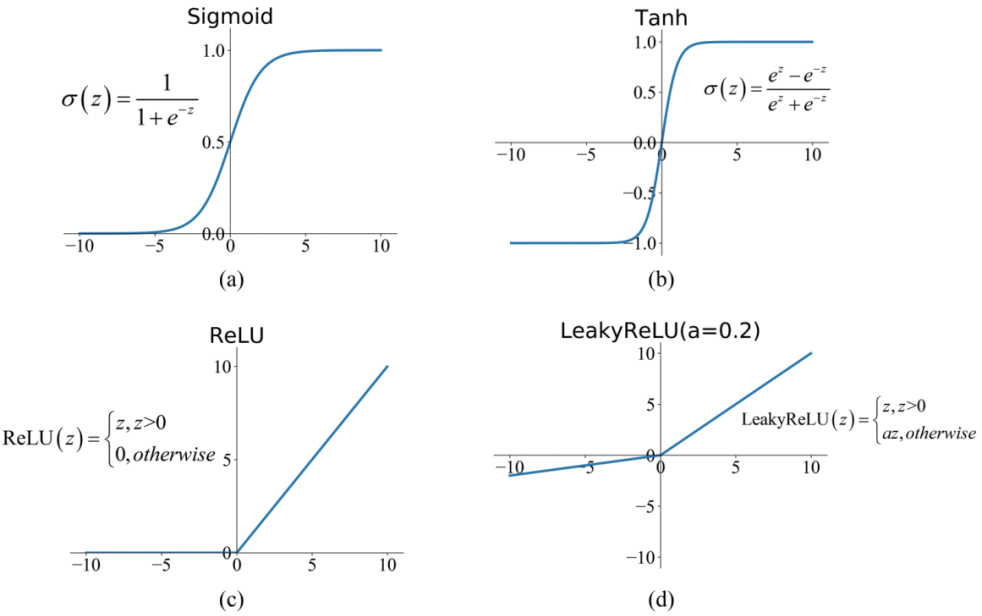
9：シグモイド関数は０～１内の値をとるため加重和の標準化が行えるほか、学習の際の誤差逆伝播法で微分値が簡便な形となるメリットがある。ReLUはシグモイド関数の弱点である勾配喪失問題（微分値がゼロに近づくと学習が進まなくなる問題）を回避するのに効果的である。

図表4　ニューラルネットワークの基本構成



出所）筆者作成

図表5　非線形変換の事例



出所）Feng et al.（2019）

金融分野の応用問題としては、デフォルト判別・与信判別分析や格付け推計モデルがある。こうした判別・分類問題に対しては、計量経済学の手法として質的選択モデル / 離散選択モデルが知られており、このほか機械学習の分野でも様々な手法が考案されてきたが、NN は関数の表現力が非常に高く、これが優れた判別能力をもたらした[10]。副島（1996）は、金融政策反応関数（当時は公定歩合の上げ / 下げ / 据え置き）に単純な構成の NN を適用し、政策変更を高い精度でインサンプル推計している。論文中では、NN 関数の柔軟で複雑な非線形性を可視化している。著名な金融政策反応関数として線形のテイラールールがあるが、これより遥かに複雑な関数形状を示しており、かつ、同じインフレ率、実質成長率に対しても、これらが加速している場合と減速している場合では、政策反応関数の形状が異なることを示している[11]。

NN は中間層の階層やインプット数を増やすとモデルの精度や表現力が高まることが理論的にも期待された。しかし、中間層を積み増して深くする（ディープにする）と学習が進まなくなるという技術的問題に直面し、80 年代の第二次 AI ブームは徐々に収束して、その後、長い AI 冬の時代が訪れた。ところが、2000 年代後半にこの限界をブレイクする技術が現れた。ディープニューラルネットワーク（DNN）であり、これが第三次 AI ブームを切り開いた。中間層を数十も積み重ねてもパラメータの学習が上手く行われる手法の開発や計算能力の確保によって、NN のパフォーマンスが飛躍的な向上を辿り始めた。LLM には NN/DNN を発展させたモデルが利用されており、これらを 3 節で解説する。

10：シグモイド関数型 NN の中間層を無くすと、質的選択モデルの代表例であるロジットモデルと一致する。シグモイド関数はロジット関数と呼称名が異なるだけであり関数形は同一である。質的選択モデルではパラメータ推計に最尤推定法が用いられるが、NN では確率的勾配降下法が誤差逆伝播法において利用される。岡崎他（2022）の第二章が両者を比較しつつ詳細に解説している。

11：副島 (1996) の図表 17、18 を参照。また、インフレ率や実質成長率の加速度が同じでも、これらの水準が異なると政策反応は当然異なってくる。

2.3 言葉の数値化：分散表現/埋め込み表現

ニューラル言語モデルの進化とあわせてLLM時代をもたらした要素に、言葉を数値化する手法、いわゆるベクトル化技術（分散表現や埋め込み表現[12]とも呼ばれる）がある。その代表例が2013年に登場したword2vecである。

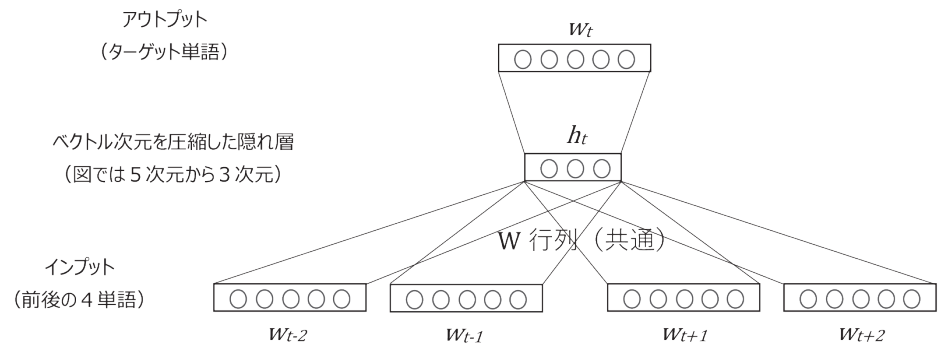
ニューラルネットワークのインプットやアウトプットは数値である。それゆえニューラル言語モデルでは、言葉を数値化する必要がある。単純なアプローチとして単語にIDを振る方法が考えられるが、膨大な量になるだけでなく、文章に内在する単語間の関係性をどのように効率的に表現するかという問題に直面する。仮に単語が100万語あって、単語1と残りの全単語との関係を個々に数字で表現していくと、その情報量は100万×100万の行列となり非常に効率が悪い。当初は、実際の文章に出てくる単語の「離接」関係（何単語分離れているか）を用いて単語をベクトル表現化するアプローチが採られた。

新しいアプローチをもたらしたのはword2vecというモデルである（Mikolov et al. 2013）。これは、文章中の前後の単語（複数単語でも可）から、間にある単語を推測する、あるいは、ある単語から前後の複数の単語を推測するという推論問題をNNに学習させるものである。前者をCBOW（Continuous Bag-of-Words）、後者をskip-gramと呼ぶ。穴埋め問題そのものに意味があるわけではなく、推定されたパラメータ（図表6におけるインプット層と中間層を結ぶ加重パラメータW）が単語間の関係性を凝縮した行列となっており、その各列が各単語の分散表現となるよう穴埋め問題が設計されている。図表6のモデルは非線形変換を含まないためNNモデルではないが、線形関係だけで表現されたモデルのため、当該単語と前後の数単語の関係をコンパクトな次元空間に押し込む/埋め込む（embed）ことができる。これが分散表現や埋め込み表現と言われる所以である。

CBOWを示した図表6・7に沿って説明する。インプットは前後の単語をone-hotベクトル表現したもの4つである。単語総数をN（例えば100万）とし、押し込む次元をhとすると、重みWは$N \times h$の行列となる。入力値が4つあるため、h次元のベクトル4つが中間層に向かって出力され、その平均値をとることで中間層の$\tilde{h}$ベクトルを作成し、これに$h \times N$のW'行列を乗じてN次元ベクトルに戻し、当該単語のone-hotベクトルを教師データとして、これにフィットするようWやW'を学習する（ただし、分散表現獲得のために必要としているのはW行列のみである）。ここではt番目の単語w_tを穴埋め問題のターゲットとしているが、このターゲットを文の冒頭近くから終点近くまで変えていくことで、学習データセットの組み合わせを多く用意できる。かつ、1文だけでなく膨大な文を対象に同様なデータセットを準備すると、どの問題にも共通して適用される固定パラメータWは、単語間の平均的な関係性を表現することになる。

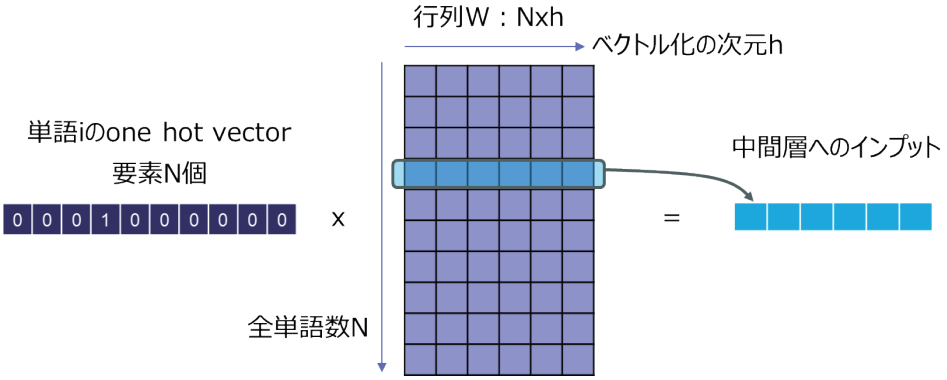
12：ある単語や文章を、高次元ベクトル空間、例えば1,536次元ベクトルに押し込んで表現するため、Embedding（埋め込み表現）と呼ばれる。Embeddingは、次元削減（高次元のデータを低次元に圧縮）や、特徴量エンジニアリング（データ特性を数値ベクトルに変換）にも用いられる。

図表6　word2vecのモデル構造



出所）筆者作成

図表7　ウエイト行列が分散表現となる仕組み



出所）筆者作成

単語をベクトル化することで以下のような言葉の計算が可能となる。

東京－日本 ＝ パリ－フランス

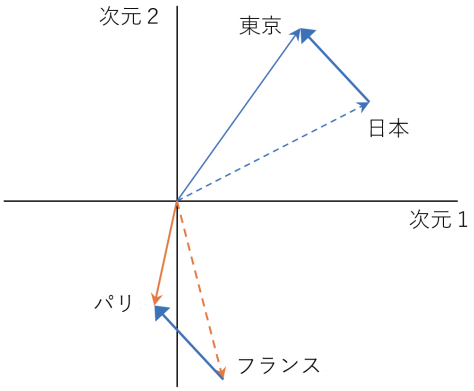
これは、概念上の情報処理ではなく、数字（ベクトル）の計算として成立しうるものである。各単語が２次元空間に埋め込まれてベクトル化されているとしよう（h=2 のケース）。このとき、あらゆる単語がこの平面上のベクトルとして表現される。日本やパリという単語が図表８に示したベクトルとして表現されたとしよう。上式左辺は図表８の太い青矢印ベクトルであり、これは右辺（オレンジで示した２つのベクトルの差）と同一となっている。青矢印は、ある国の国名と首都という単語間の関係性を示すベクトル表現に相当する。学習対象のコーパスに「日本の首都は東京である」「フランスの首都はパリである」といった単語の関係性を示す文章が数多く存在し、その関係性を word2vec が集約するため、４つの単語の関係性を分散表現で表すことが可能となる。

ベクトル化により単語の類似性の計測も図表９のように可能となる。単語が数値化されているのでコサイン類似度や空間内での近傍関係を測る指標を活

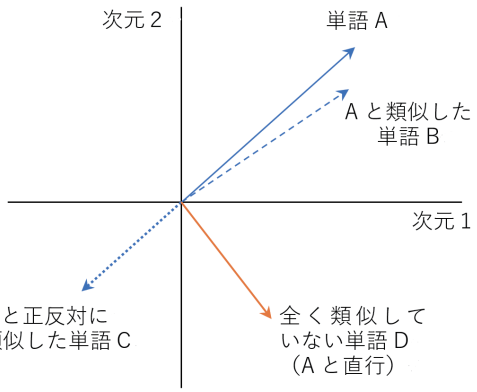
用することができる。EC や映像音楽の配信などで各種の推薦サービスが提供されているが、こうした推薦サービスには画像や商品特性を特徴量抽出し、デジタル情報化（ベクトル化）したうえで、類似度を計算するという技術が活用されている。

コサイン類似度はベクトルの内積表現から得られている。内積は 2 つのベクトルが成す角が 0 度に近い（同じ方向を向いている）ほど大きくなり、直行するとゼロとなる。内積が類似度や関連性の高さを示すことは、後述の Transformer モデルの Attention 機構で活躍するだけでなく、ニューラル言語モデル全般において重要な役割を果たしている。

図表8　パリ=フランス+（東京−日本）



図表9　二次元空間で測った単語の類似性



出所）筆者作成

次元数 h は様々な設定が可能であるが、単語間の関係性を表現するためには数百～数千次元のものが使われることが多い。N 個（仮に 100 万個）の単語間の関係を示すことに 100 万× 100 万の行列を使うことは不効率である。N が 100 万だとしても、埋め込み表現を使うことにより次元空間 100 万× h まで圧縮することが可能となる。単語の分散表現は、単に言葉を数値化するだけでなく、こうした次元圧縮によって LLM の性能向上や計算コスト削減をもたらしている。機械学習の特徴量表現においても膨大なデータ内にある関係性を同様な次元圧縮技術を用いて抽出しており、コンセプトは同じである。なお、ある分散表現モデルで埋め込み次元数 h を決定すると、どのような単語であっても固定長 h 次元のベクトルに変換される[13]。

こうして word2vec は注目を集めることとなったが、学習対象のコーパスに含まれる単語数が 100 万のように大きなものになると計算量は膨大なものとなる。Negative sampling のような効率化手法が考案されたが限界があった。また、単語の隣接関係に基づく推論問題から算出されるため、位置関係より更に複雑な文脈の流れや、複数の意味や読み方を持つ文脈依存の多義語を扱うのが不得意であった。なによりも、word2vec は言語を生成するモデルではなく、単語を分散表現する手法であり、生成 AI の実用化はその後の LLM の発展があってのことであった。

なお、分散表現のモデルは言語生成モデルと並行して発展を続けており、有名なものとしては FastText（Bojanowski et al. 2016）や ELMo（Peters

13：英語のように単語間にスペースで区切りがある、すなわち分かち書き（tokenization）がなされている言語の場合、分散表現によって各単語がベクトル化される。一方、日本語のように分かち書きしない言語では、その言語専用に開発された形態素解析で品詞に分解する前処理作業を行うことが伝統的な自然言語処理において一般的である。こうしたトークナイゼーションを行うツール（トークナイザー）としては MeCab や ChaSen、JUMAN、Janome などが知られている。しかし、分散表現モデルを適用すると、形態素解析とはかなり異なる切り方、人間には解釈が困難な切り方をする場合が少なくない。言葉のパターン性や単語・文字レベルでの関係性の把握が、人間の認知とは異なる方法で行われていると考えられる。本稿では単純化のために、単語単位にトークン化されるという想定で説明を行っている。モデル解説で単語と呼称している箇所のなかには正確にはトークンと呼ぶべきケースがあるが、わかりやすさの観点から単語と表記している。

et al. 2018）がある[14]。ELMo（Embeddings from Language Models）では双方向 LSTM（LLM の前駆的モデルの一つ、後述）を用いて、文章全体の情報に基づき単語を分散表現し（contextualized word embedding）、かつ１つの単語に３つのベクトル（各々 1024 次元）をもたせることで多義語対応などのパフォーマンスを改善している。word2vec のように前後の数単語との関連性情報だけでなく、LLM を使うことによって文脈を反映させた分散表現となっている点が大きな改善点である[15]。

ChatGPT で知られる OpenAI もベクトル表現に特化した LLM を開発している。例えば、GPT-3 や 4 では、text-embedding-ada-002 という分散表現に特化した軽量高速のモデルが用いられている[16]。ただし、GPT など近年主流となっている LLM の基礎となった Transformer モデル（Vaswani et al. 2017）では、分散表現の過程を分離するのではなく LLM の内部に取り込む手法が登場している。初期の Transformer モデルである BERT（Devlin et al. 2018）がこうした流れを作った[17]。この点は 3.4 節で後述する。にもかかわらず、分散表現モデルの開発が続いているのは、RAG でベクトル化データベースを作成するニーズ（後述）や、運用の効率性追求、計算資源の節約といったが理由が存在しているためである。

３．LLM時代をもたらしたニューラル言語モデル

LLM の発展には、単語が意味ある順序で並んでいるという文章の系列データ的特性を表現できるモデルが必要であった。株価の時系列データを並び替えると意味がなくなるのと同様、単語も文書内でランダムに並び替えることはできない。その意味で、文書は単語の時系列データである。当初のニューラル言語モデルでは、t 番目の単語をアウトプットとして生成するために、直前の n 個の単語をインプットデータとしていた。これは、単語や文脈の繋がりを反映させるために必要な処置であった。しかし、計算負荷の観点から扱える語彙数を絞る必要があり、また n 個の単語の前後関係情報も活用されなかった。

これに対し、既存の RNN モデル（Recurrent NN）を自然言語処理に応用しようという試みが登場した。RNN では、過去にどのようなインプットがあったかという情報を累積・更新しながら引き継いでいく状態変数ベクトルが導入されている。また、同モデルの欠点を補う LSTM モデル（Long Short Term Memory）も開発されていたので、RNN と並行して LSTM の応用も進められた。しかし、両モデルとも状態変数を導入したことにより、NN のパラメータ算出に用いるバックプロパゲーションが関数の入れ子構造をとることになった。これが計算負荷や学習における難点となっていた。また、インプットとして長い文章を扱うと、入れ子構造が長大化するため、長文を処理することが苦手であった。この問題を解決したのが Transformer モデルの Attention 機構である。３節では、これらのモデルを順に解説する。なお、本節の内容は、各モデルを提唱した原論文に加え、岡崎他（2022）、Rothman（2021）、坪井他（2017）、斎藤（2018）を参考にしている。

14：これらのほかにも、word2vec と同様な発想で文章を分散表現するモデル Doc2Vec が考案された（Le and Mikolov 2014）。しかし、Transformer モデルで LLM の内部に分散表現を取り込んでいく動きが主流になったことや、対象が文章全体であったことから（単語単位、文単位、数文単位など区切り方を自在にできる分散表現のほうが便利である）、発展はしなかった。

15：文中の単語をマスクし予測させることに加え、文の前後関係の正誤判別トレーニングという２つの学習手法を適用することで、文脈を捉える能力を高めている。単語レベルでの分散表現獲得では、word2vec は前後の単語の的確な選択訓練に skip-gram を用いているが、これを文単位版で学習する発想に類似している（正確には、文そのものを予測はせず、順番をバラバラにした候補文＜教師データ＞の中から適したものを選択するという学習である）。

16：2023 年末には、text-embedding-3-small/large にモデルチェンジしている。

17：これに対し、ELMo は双方向 LSTM という LLM を使いつつも分散表現を作成することが主眼となっており、それ自体に文章生成能力はない。BERT は LLM の機能のなかで分散表現を行っており、主たる用途はエンコーダー機能（文書分類等）である。

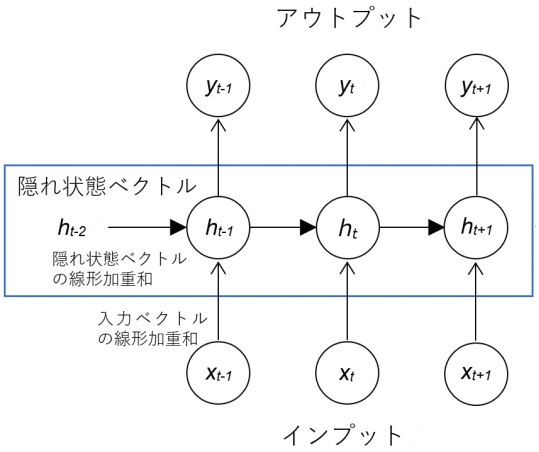
3.1 RNN：系列データへの対応

1980 年代央に登場した初期のフィードフォワード型 NN はインプットの順序性を考慮しなかった。どのような順番で並べても、対応する重みパラメータ群が入れ替わることで対応できたが、見方を変えると、株価のような順序性がある時系列データや、文章における単語のような系列データにおいて、並び方が持つ情報を捉えることができなかった。この問題に対応するため、NN が登場して間もないころから、時間展開や前後展開の情報を取り扱える RNN モデルが Jordan（1986）や Elman（1990）によって提案されていた。

RNN では過去にどのようなインプットがあったかという情報を累積・更新しながら引き継いでいく変数が導入された。これを隠れ状態ベクトル h_t と呼ぶ。図表 10 のように一期前の隠れ状態ベクトル h_{t-1} と今期のインプットベクトル x_t を合算することで、隠れ状態をインプットごとに更新していく。Recurrent は再帰という意味だが、隠れ層の情報がアウトプットに向かうだけでなく、内部（翌期の隠れ層）に再帰するため、こうした用語法が用いられる[18]。なお、時系列モデルの視点からは、RNN は状態空間モデル（カルマンフィルター）の NN 版と捉えることができる。

18：図表 4 で示したフィードフォワード型の NN とは別に、Hopfield（1982）は全要素が結合したホップフィールドネットワークを提唱している。これは、出力がネットワーク内に再帰（recurrent）するという点で RNN の前駆的なモデルとなっている。

図表10　RNNの基本モデル



出所）筆者作成

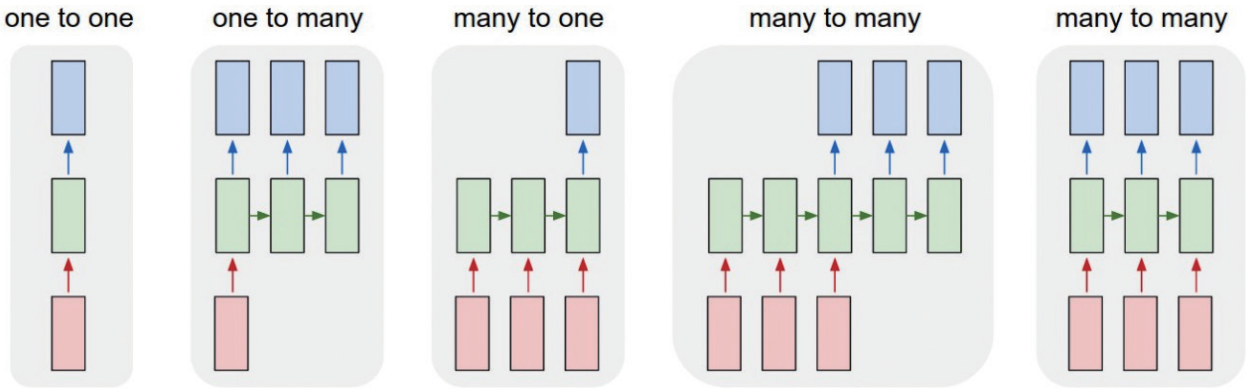
RNN は時系列データを扱えるモデルとして応用が進んだが、この時期はニューラルネットワークの学習能力の限界や問題点が判明した時期と重なり、その後、いわゆる AI 冬の時代を迎えることになった。このため RNN モデルが注目される機会は減っていった。ところが、自然言語処理の研究が進展するなかで、2000 年代初頭にニューラルネットを自然言語処理に応用することで文章を生成するモデル（ニューラル言語モデル）を作る試みが Bengio et al.（2000）や Bengio et al.（2003）により始まった。そして、Mikolov et al.（2010）によって RNN を言語モデルに応用した研究が登場し、RNN は再び注目を集めることになった。

図表 10 では、各インプットや一期前の隠れ状態から出ている矢印が一本で表現されているが、分散表現によって各単語が高次元ベクトルに変換されているため、実際には多くの数値を加重和として集約している。隠れ状態も高次元ベクトルであるため、こちらも加重和したうえで、インプット側と合算される。これにバイアス項（定数項）を加えたうえで非線形変換し、アウトプット層へ引き渡される。なお、線形加重和や非線形変換のパラメータは、各 t 期において共通しており、インプット内容に応じて変化したりはしない。言い換えると、どのようなインプットや隠れ状態ベクトルに対しても共通して対応する高い汎用性を持つ。

隠れ状態ベクトルは、過去のインプットを順次集約して積み重ねていったものであるため、言葉の並びが作り出す意味・文脈情報が埋め込まれている。こうした RNN の仕組みにより、文章に対応する能力が高まった。図表 10 では、隠れ状態ベクトルは 1 層となっているが、複数の隠れ状態ベクトルを階層構造として導入することができる。これは、ディープニューラルネットワークのアプローチと同様である。図表 4 に示した基本形のフィードフォワード NN では、中間層を複数階層に積み上げる（ディープにする）ことで、インプットとアウトプットの間にある複雑な情報構造を捉える能力を高めることができる。言い換えると、インプットをアウトプットに変換する関数として、より複雑で高性能なものを作り出すことができる。RNN の隠れ層にも同じコンセプトを適用し、深く階層化することで能力向上を図ることが可能である。

次に、RNN には様々な応用バリエーションがあることを紹介する。図表 10 では、インプットとアウトプットが一対一に対応していたが、様々なバリエーションが設計可能である。代表的な設計例を図表 11 に示した。一番左は、インプットとアウトプットが 1 対 1 に対応している事例である。最もシンプルなこのケースでは、中間層があるものの、その内部の隠れ状態ベクトルが横方向（時間展開方向）に引き継がれていないため、RNN とはいえない。説明のための基本形として取り上げたものである。しかし、このシンプルなケースも用途は多々あり、例えば、画像認識や分類・識別問題がこれに相当する。 1 つの画像をインプットとし、それが何を示しているかをアウトプットとした画像の分類・識別の事例である。

図表11　RNNの様々な設計例



出典）Karpathy (2015)

左から２番目は、画像からの文章生成事例である。文章を単語の逐次生成によって作っていくため、アウトプットが順次作成されている。一つの画像インプットから最初の単語を生成し、隠れ状態ベクトルを更新しながら、その単語が生成されたもとで次の単語を選択していく手続きをとる。

３番目は、文章からの内容分類などが相当する。アウトプットは内容に関する複数の分類候補であり、逐次入力されるインプットを最後まで受けたうえで、アウトプットが生成される。例えば、この文書のトピックは何か（国際政治、経済、スポーツ、芸能、金融市場等）を識別させる問題である。企業の決算報告を読み込んだうえで、今後の株価が、上がる・下がる・横ばいという離散選択肢を選ばせる問題もこれに相当する。

４番目は翻訳が典型例である。日本語の文章をインプットとし、最後まで読み込んだところで、英語の文章を順次生成していくケースに相当する。５番目は、同時通訳やビデオ画像へのキャプション文書生成などに相当する。４番目との違いは、インプットの読み込みを最後まで待たずに、順次アウトプットを生成していく点である。

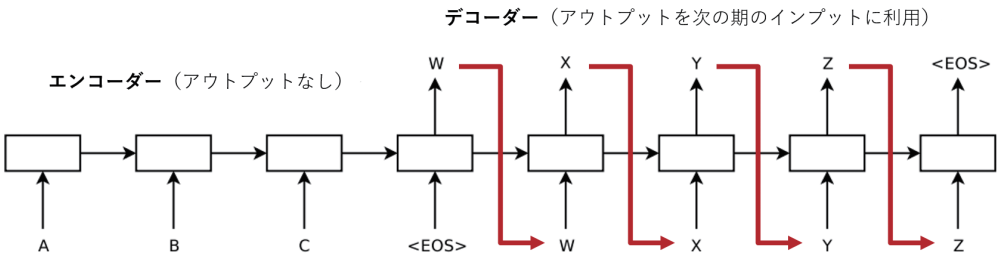
3.2 エンコーダー・デコーダーモデル、seq2seqモデル

上記のようにRNNの応用モデルが広がるなか、RNNを２つセットで利用するseq2seqモデルがSutskever, Vinyals and Le（2014）によって提案され、機械翻訳などへの応用が加速した（正確にはRNNの発展形であるLSTMを利用している）。図表12に示したようにseq2seqでは一つのRNNでインプットとアウトプットを対応させるのではなく、前半のRNNでインプット情報を隠れ状態ベクトルに集約させ、これを後半のRNNの隠れ状態ベクトルの初期値として引き渡し、アウトプットを順次生成していく。後半のRNNのインプットには、一期前のアウトプットが用いられている。例えば日英翻訳の場合、日本語文章を読み取った後、生成された英単語１つが逐次的に翌期のインプットとなり、隠れ状態ベクトルと合わせて翌期のアウトプット（次の英単語）を生成する。隠れ状態ベクトルには前半で読み取った日本語

の情報が蓄積されており、英単語を一つ生成するたびにアップデイトされていく。

単語の並びとしての文章内容を高次元の固定ベクトル（ここでは隠れ状態ベクトル h_t）に変換することを、エンコーディング（符号化）と呼び、これを順次、文章に復元することをデコーディング（復号化）と呼ぶ。2 つの RNN をエンコーディングとデコーディングに特化させて使うアプローチであり、のちに LLM の性能を劇的に改善させた Transformer モデルも、エンコーダー・デコーダーモデルとして作成されている。エンコーダー部分はアウトプットを持たず、デコーダー部分にアウトプットを生成させることで 2 つの RNN がセットで学習されることになる。通常の RNN では各期の線形加重和や非線形変換のパラメータは共通となっているが、seq2seq ではエンコーダー部とデコーダー部の RNN でパラメータが異なる分、モデルの自由度が高まり、学習性能も向上する。

図表12　seq2seqモデル



出典）Sutskever et al.（2014）
注）EOSは文章の終了を示す。オリジナル論文の図表に筆者が加筆。

なお、seq2seq モデルの用途は機械翻訳に限らない。文章を与えて要約文を返す、質問を与えて回答を返す、対話を与えて対話を返す、文章を与えて文法誤り修正後の文章を返す、口語文章を与えてビジネス文章に変換して返すなど、文字系列を異なる文字系列に変換する様々な用途がある。

seq2seq モデルは、その名前の意味からすると、文章のような系列インプットを系列アウトプットに変換するモデルを指すことになる。ほぼ同時期に同様なモデルを開発した Cho et al.（2014a）は、seq2seq ではなくエンコーダー・デコーダーモデルという呼称を用いている。3.4 節で示すように Transformer はその名前の通り、機能的には系列変換モデルであり、モデル構造としてはエンコーダー・デコーダーモデルを採用している。Transformer 系統のモデルが主流となった近年においては、汎用アーキテクチャとしてのモデル構造に注目してエンコーダー・デコーダーモデルという呼称が使われることが多い。

こうして登場から十数年の時をおいて、時系列データから自然言語処理に用途が拡大した RNN だが、モデルの構造上、2 つの課題を抱えていた。一つは、隠れ状態ベクトルのアップデイトをどれほど強く行うか、すなわち、新し

いインプットの情報を強く反映させるか、それとも過去のインプット情報を長く保有させるかというトレードオフ問題である。前者の場合、過去の情報が損なわれやすく、文脈を長く保持しておくことができない。逆に後者の場合、新しい単語がもたらす文章内容の展開について行けなくなるという問題がある。適切な単語の選択は直前の単語の影響を強く受けるため、直前や近い位置のインプット情報を重視せざるをえない。このため、RNN では遠く離れた単語や文脈をうけて次の単語を適切に選択することが難しかった。言い換えると、隠れ状態ベクトルにおいて長期記憶を保持することが困難であった。

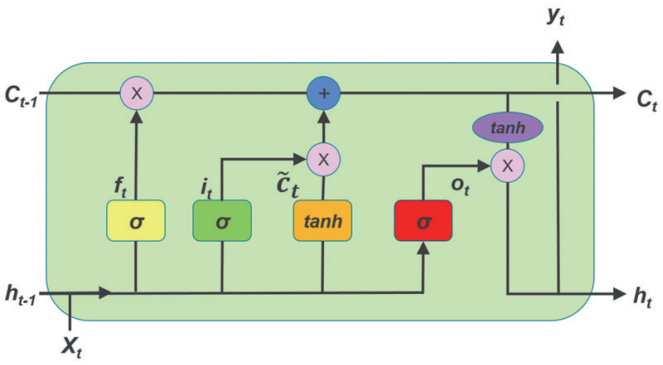
RNN が自然言語処理に適用される前から、こうした問題は認知されており、解決する方法が考案されてきた。1990 年代後半に登場してきた LSTM（Long Short Term Memory）が代表例である。2010 年代になると RNN と同様に自然言語処理への LSTM 適用が進展した。

なお、RNN が抱える 2 つの課題のうちのもう一つは、学習に誤差逆伝播法（の応用系）を適用する際に、計算が不安定化もしくは進行しなくなり精度が上がらなくなるという学習時の難点であった。この問題は LSTM でも解決できておらず、Transformer モデルの登場を待つことになる。

3.3 LSTM：短期記憶と長期記憶

Hochreiter and Schmidhuber（1997）は、前述のような RNN の課題について長期記憶と短期記憶に役割分担させることで解決を図った。図表 13 は隠れ層の内部を示している。h_t は RNN と同様、隠れ状態ベクトルを示すが、その役割は短期記憶に特化している。長期記憶は c_t が担っており、左上の前期値から引き継がれてきた c_{t-1} は、まず忘却関数 f_t を通過する。忘却の程度の調整は、短期記憶 h_{t-1} と今期のインプット x_t の線形和にシグモイド関数 σ を適用することで0～1の範囲で調整される。この仕組みは忘却ゲートと呼ばれている。次に新規情報を長期記憶に追加する。新規情報も h_{t-1} と x_t から作成され、別の加重和とハイパーボリック・タンジェント（tanh）関数によって作成されたベクトル情報（図中ではセルゲートを通過する $\tilde{c}_t$、g_t と表記する文献もある）に、入力ゲートの重み（入力ゲート関数 i_t で作成された0～1内の値）を乗じたものを加算する。この重みも h_{t-1} と x_t から作成される。これを翌期に引き渡す（図中の右上に抜けていく c_t）ほか、短期記憶 h_t は長期記憶 c_t の一部を残すことによって作成される。まず、c_t に tanh 関数をかけて基準化し、これに出力ゲート o_t からの0～1調整を乗じて、今期の短期記憶 h_t とするとともに、これをアウトプット y_t として利用する。出力ゲート o_t も h_{t-1} と x_t から作成され、シグモイド関数によって範囲調整される。

図表13　LSTMの短期記憶と長期記憶のアップデイト



出所）Zheng, Yuan and Chen（2017）

RNN では中間層に相当する h_t は更に加重和と非線形変換を通じてアウトプットに変換されていたが、LSTM では短期記憶をアップデイトしたものをそのままアウトプットとしている。今期の短期記憶は、今期の長期記憶を tanh 関数で基準化したものが元情報になっており、２つの状態変数ベクトルがアウトプットに近い情報セットとなっている。見方を変えると、長期記憶の更新が RNN における隠れ層からアウトプット層への変換に相当していると解釈することもできる。こうした設計は LSTM の計算効率を高めるうえでの工夫として導入されている。

LSTM のモデルの特徴の一つは、あらゆる情報が短期記憶 h_{t-1} の加重和と今期インプット x_t の加重和の和で更新されていく点である。忘却ゲート、入力ゲート、出力ゲートを作成するパラメータ群は全て学習によって算出され、これらの違いが各種ゲートの機能の相違や長期記憶に付加する内容そのものを生み出している。Transformer モデルの Attention 機構（QKV 機構、後述）では、こうした発想が更に大規模に活用される。

RNN が自然言語処理に適用され始めたのは 2010 年であったが、1997 年に登場していた LSTM も同時期に自然言語処理に応用され始める。前述したように 2014 年には seq2seq モデルに LSTM を適用した研究が登場している（Sutskever, Vinyals and Le 2014）。なお、SLTM と類似したモデルとして GRU（ゲート付き再帰ユニット）がある。本稿では説明を省略するが、そのエッセンスは、長期記憶は導入せずに状態変数ベクトル h_t についてリセットゲートや更新ゲートを通じて更新していくというものである。GRU は最初から自然言語への応用を念頭に開発されたものである（Cho et al. 2014b）。Cho らは、同年にエンコーダー・デコーダーモデルの出発点となった研究も行っている（前述の Cho et al. 2014a）。

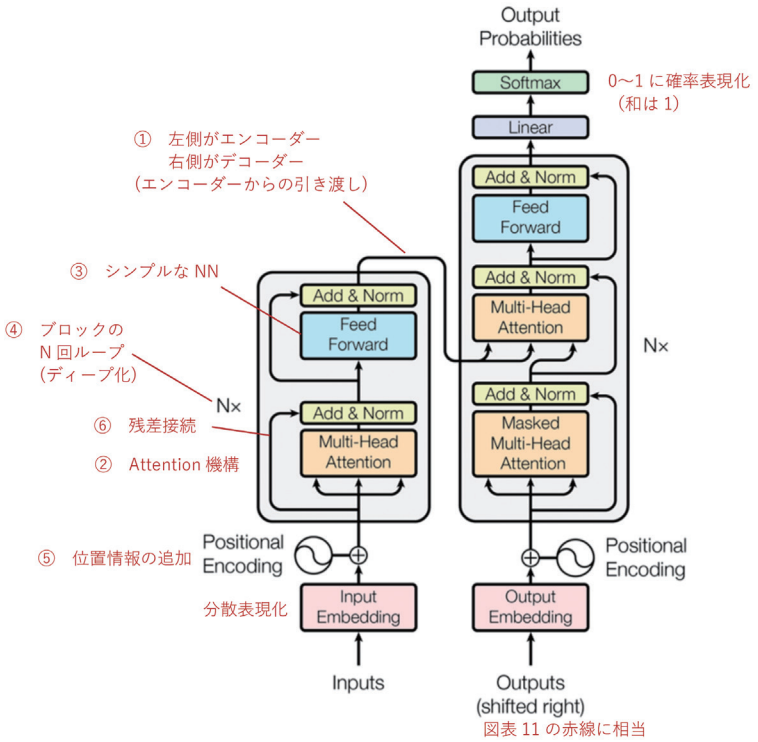
３.４ Transformerモデル：Attention機構によるゲームチェンジ

RNN、LSTM、GRU モデルは、いずれも再帰モデルであった。モデルが時間方向・系列方向に入れ子構造となっている点は情報更新の面で機能的であるが、学習の際の計算負荷が高く、かつ学習が進みにくい（勾配消失問題や勾配爆発問題）。加えて、学習後の推論（インプットを与えてアウトプットを得

る）の際にも逐次的処理となり、言語モデルとして利用する際に、並列処理による時間短縮ができないという問題があった。入れ子構造は長文処理能力の強化においても制約となり、長文処理の制約によって複雑な文脈構造や指示関係も捉えにくくなるという限界もあった。

これらの点を改善したのが、Attention 機構を持った Transformer モデルである。“Attention is all you need”というタイトルが付けられた Vaswani et al.（2017）は、現在の LLM 大進化の起点となった。同モデルは、図表 14 で示しているように、①エンコーダー・デコーダーモデルの適用、② Self-Attention、Malti-head Attention 機構を取り入れて、単語の重層的な文中参照状況を検知・活用、③ Attention 機構が関連付けた情報をフィードフォワード NN で処理、④これらから成るブロックを多数重ね上げることで、序盤のブロックでは表記的な、中盤では文法的な、終盤では意味的な処理を多重実装（同論文では 6 回ループする）、⑤インプットの分散表現に文章中の単語の位置情報も追加し、前後関係の情報を強化、⑥残差接続[19] によって勾配消失問題の緩和や位置情報の保持（Attention 機構を通る際に情報がロストする点を迂回補強）を達成、といった特徴を有している（番号は図表 14 中の付番箇所に相当）。

図表14　Transformerモデルの構成概要



出所）Vaswani et al.（2017）
注）現論文の図表に筆者が補記（赤字赤線部分）

以下では、Attention 機構のコアとなる Scaled Dot-Product Attention について解説する[20]（Dot-Product は内積、Scaled は基準化）。図表 15 に計算過程を示している。

19：残差接続 / 残差結合（residual connection）とは、ひとつ前の層の結果を現在の層の出力結果に足し合わせる処理である。層 i への入力値を P_i、出力を O_i とすると、$P_i=O_i+P_{i-1}$ となる。出力結果に元の入力値を変形すると $P_i=O_i+O_{i-1}+O_{i-2}+O_{i-3}+\cdots+O_0$ となり、下層の出力を加算していったものが上層への入力値となる。残差という表現は、出力値が入力値の差となる点（$O_i=P_i-P_{i-1}$）から生じていると思われる。残差接続は、勾配消失問題の軽減のほか学習の安定化（一度に大きく変化させない）という効果を有する。

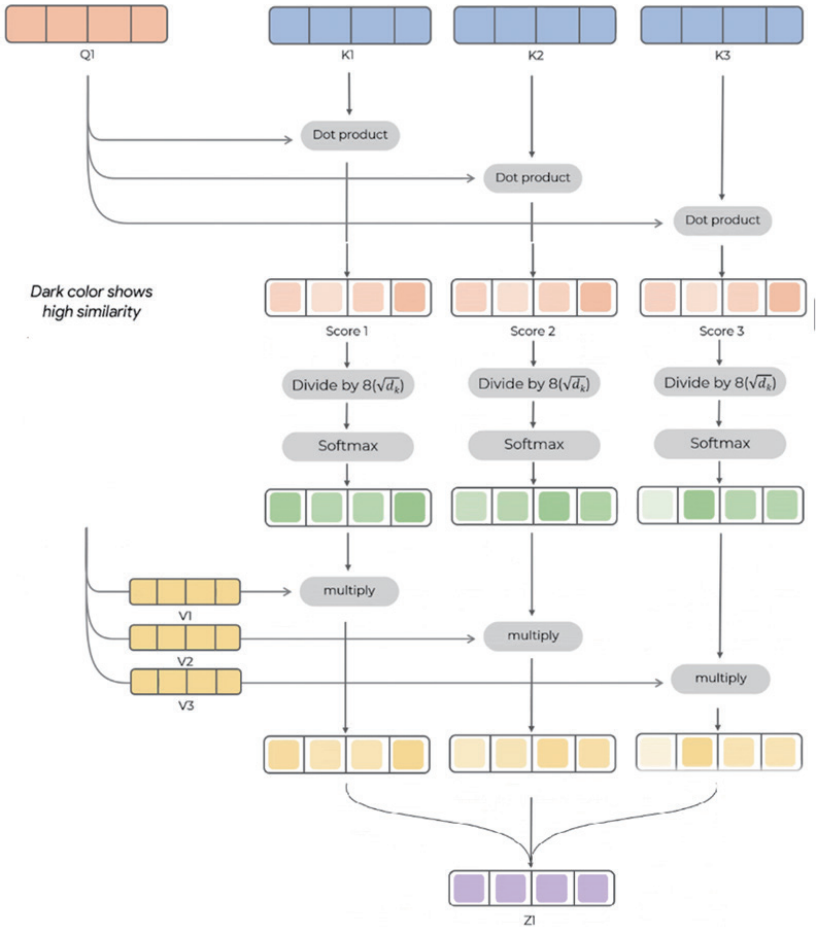
20：Attention 機構のアイディアは Transformer モデルが登場する前から Bahdanau, Cho and Bengio（2014）によって提唱されていたが、注目を集めたのは Transformer モデルに Scaled Dot-Product Attention の QKV 機構として組み込まれてからである。一般に Attention 機構という場合は、Transformer モデルに組み込まれた Scaled Dot-Product Attention と Multi-head Attention を指す。

まず、クエリ（Q）、キー（K）、バリュー（V）という概念を導入する。クエリ（問い合わせ）は、ある単語を対象にしたとき、文中の他の単語に対してどのような関連性があるかを調べる際の問い合わせ内容である。キーは、この問い合わせを受ける内容（関連性が評価される対象内容）である。クエリとキーの内積を計算することで関連性の強さを計測する（内積が類似性を示していることは 2.3 節の word2vec を参照）。これに softmax 関数を適用することで、内積をとった結果を確率表現化する。softmax 関数は、和が 1 となるよう入力値を変換するものであり、NN 全般においてアウトプットを確率として表現する際に用いられる（2.1 節で触れた「東京」が選ばれる確率や他の単語が選ばれる確率を計算する際にも同関数が利用されている）。この確率をバリューに対して乗じる。バリューは、インプットを分散表現（位置情報エンコーディングを加算）したベクトルの加重和であり、この中のどれに重きを置いて取り出すかを、Q と K の内積を通じて得た関連性情報に基づいて決定する。これが、Q、K、V 概念を使った Scaled Dot-Product Attention の仕組みである（図表 15 参照）。

ちなみに、Q、K、V のどの要素もインプットの分散表現をそのまま使うのではなく、その加重和を利用する。いずれもが同じ情報セットから加重和の違いによって作られるという発想は、LSTM の解説で示した全ての情報が h_{t-1} と x_t の加重和で更新されていくという NN の一般的な発想（全情報の中にある何らかのパターン性を学習によって発見して活用する）と同じである。一つ前のパラグラフで、バリューは、インプットの分散表現ベクトル「の加重和」と表記しているが、これは、問い合わせ（クエリ）との関連性を評価するキーと、キーの内容を示すバリュー、これらに出てくる Q、K、V の全てをインプットの線形結合で表現している仕組みが採用されているためである。インプット文章を構成する単語列という文書の内部構造に対して自己調査的に Attention を計測するため、Self-attention と呼称されている。

なお、デコーダー部分の下位の Attention 機構では、デコーダーが生成したアウトプットを順次利用するが、その先のアウトプット部分が先読み参照されないよう情報が遮断（mask）されている。また、上位の Attention 機構では、QKV として三本の矢印が異なる箇所から入力されている。エンコーダーのアウトプットからキーとバリューの元となる情報を受け取り、下層の Attention 機構のアウトプットからはクエリの元となる情報を受け取っている。これは、デコーダーにおいて作成済みのアウトプット（単語）が、次にくる単語の適切な選択を考える際のクエリ（問い合わせ）起点になるためである。それが指す対象であるキーや内容に相当する V はエンコーダーで抽出された情報を参照している。これが、上位の Attention 機構の QKV の特徴となっている。図表 12 の seq2seq モデルでは、デコーダーの一期前アウトプットを今期のインプットとし、エンコーダーから受け継いだ状態ベクトルをアップデイトしていっているが、これをより複雑に設計したエンコーダー・デコーダーモデルとなっている。

図表15　Scaled Dot-Product AttentionにおけるQKV計算



出所）Harsoor（2023）

注）QとKの内積に対して次元数d_k（ここでは64）のルートをとったもの、すなわち８で除算をしているのは、次元数が高いほど内積値が大きくなるため、その効果を基準化したものである。オリジナルのTransformer論文で導入されている。なお、Transformerモデルの動き方をExcelで追ってみることができる研究サイトがある。GPT-2を小規模にしてAttention機構の動きを数値例でモニターすることができる。

Ishan Anand, “Spreadsheets are all you need,” https://spreadsheets-are-all-you-need.ai/, (GitHub) https://github.com/ianand/spreadsheets-are-all-you-need/

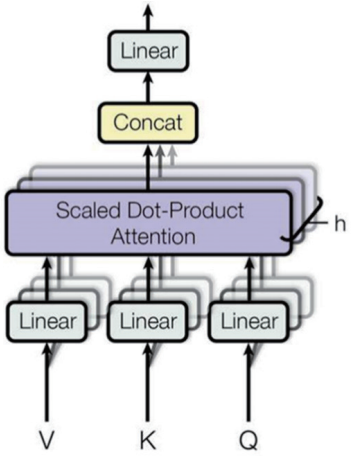
ところで、文中のある単語が指す、あるいは強い関連性を持つ単語は複数ある。関連性を捉える視点も複数存在する。これらに対応するためにはAttention機構を複数用意する必要がある。Transformerモデルを提唱したVaswani et al.（2017）では、インプットの各単語を512次元ベクトルに分散表現しているが、これを８つのグループに分けて、各64次元のベクトルに対しScaled Dot-Product Attentionを適用することで、複数の関係性を把握しようとしている[21]。これがマルチヘッドアテンション（Multi-head Attention）の仕組みとなっている。図表16にはh個のAttentionの結果が、単純に連結（concat）されて、更に重みづけ調整（図中の最上位のLinear）される仕組みが示されている。これが図表14中の残差接続と正規化作業の直前の箇所への出力となる。

21：８グループへの切り方は512次元の前方から順に切っていくだけであるが、そのように分断された分散表現に対して個々にAttention機構を適用するとAttentionのパフォーマンスが低下すると考えられる。しかし、実際にはTransformerモデルの性能は大きく改善しており、関係性の発見を多様な視点で与えることのメリットが上回っているものと思われる。

図表 17 にはマルチヘッドアテンションが実際にどのように単語間の参照関係を計測しているか調査した事例を紹介している。左側では、名詞句に形容詞句がどのように修飾関係をもってかかってくるかを探知する Attention が作成されている。右側は受け身関係にある be 動詞＋過去分詞の関係を探知する Attention が作成されている。こうした様々な Attention はモデルの学習（NN における学習）によって獲得される。

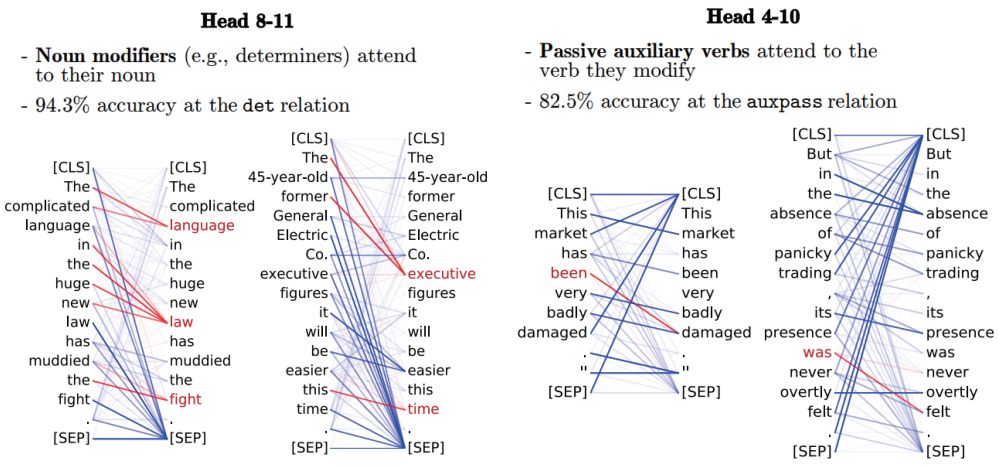
Transformer モデルは、RNN のような入れ子構造で単語間の関係を調べるのではなく、Attention 機構を活用するため並列処理が可能となり、計算負荷の軽減やモデルの大規模化に繋がった。マルチヘッドアテンションも、分散表現によるベクトル次元サイズを分割することで並列処理を更に強化している。

図表16　マルチヘッドアテンション



出所）Vaswani et al.（2017）

図表17　マルチヘッドアテンションの事例



出所）Clark et al.（2019）

Transformer モデルは Attention 機構に注目が集まりがちだが、単語間の関連性や指示関係性が組み込まれた状態ベクトルは、フィードフォワード NN のパラメータ群で整理される。言い換えると、学習された情報はこのパラメータ群に蓄積され、知識となる。Attention 機構を含むエンコーダー側のブロック（図表 14 のグレー部分）は 6 回積み重なったうえで、デコーダー側に情報を引き渡すが、上位階層のフィードフォワード NN ほど高次な意味概念を蓄積する役割を果たしている。また、同 NN は、インプットの次元が 512 であるのに対し、中間層が 2048 と 4 倍増しており、特徴量抽出の機能が強化されている（ただし中間層は 1 つのみである）。

3.5 LLMの群雄割拠

Transformer モデルの登場は、ニューラル言語モデルの能力を劇的に高めた。最初に注目を集めたのが Google の研究チーム、Devlin et al.（2018）によって考案された BERT（Bidirectional Encoder Representations from Transformers）である。BERT や同年登場した GPT は次節で触れる事前学習（Pre-training）とファインチューニングのパラダイムを確立し、その後の LLM の発展方向に大きな影響を与えた。また、BERT は Transformer モデルのうちデコーダー部を利用せず、エンコーダー部のみを用いた点も特徴的である。その代わり、分散表現の次元数が 768、ブロックのループが 12 回もしくは 24 回（BERT Large モデル）、マルチヘッド数は 12 もしくは 16 とオリジナルの Transformer モデルより規模が大きくなっている。BERT の成功により、デコーダー系のモデルとして RoBERTa、DistilBERT、ALBERT、XLNet など様々なバリエーションや改良版が登場した。

同 2018 年には、その後の生成 AI/LLM ブームを引き起こした GPT（Generative Pre-trained Transformer）の初代モデルが OpenAI 社によって開発された（論文公表時でみると GPT が先行している）。BERT と同様、その名前には Transformer が含まれている。2010 年代を通じてニューラル言語モデルの開発拠点が大学から企業へシフトしつつあったが、この時期になると、開発の主導がプラットフォーマー大企業や IT 企業の研究チーム中心となる[22]。そうした LLM モデル群が登場した 2018 年から 19 年にかけては、モデルの大規模化が起こり始めた（更に加速したのはその後であり、次節で解説する）。LLM（大規模言語モデル）という言葉は、Transformer モデル登場後のモデルの大規模化とこれに伴う性能向上によって誕生し、実務への応用が進んだことによって一般に知られていくことになる。

図表 18 は、そうした LLM の進化を樹形系統図で示している。2023 年初までの状況であり、例えば Google の PaLM2 を引き継いだ Gemini は図中には登場していないが、Transformer 以降の LLM 時代の始まりを概観するのに有益である。同図は、2018 年の段階で Transformer モデルが 3 つに分岐したことを示している。エンコーダーモデル（樹形図の左側）、デコーダーモデル（右側）、エンコーダー・デコーダーモデル（中央）である。BERT がエンコーダーの代表例であり、2019 ～ 20 年にかけて派生拡張モデルを生んだが、その後は続いていない。LLM 開発においては文章理解に加えて文

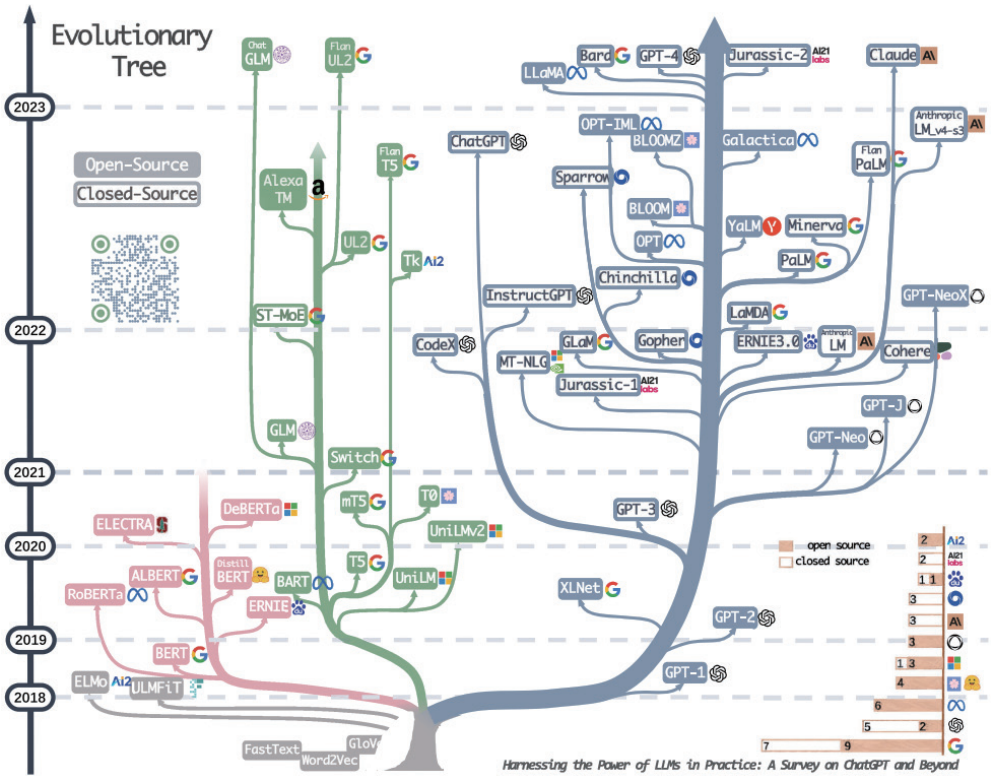
22：本稿の参考文献を Google scholar 等で検索すると論文を参照できる。表紙の筆者所属情報を時代順にみていくと、こうしたシフトが観察される。また、共著者数の増加も企業主導型になったことの反映であると思われる。執筆者が 10 名を超える論文も少なくない。

章を生成する機能も重視され、また、画像や音声の生成などマルチモーダル化が進んだため、復号機能を持つデコーダー系に数多くの LLM が登場した。OpenAI の GPT シリーズもデコーダー系である。そもそも、デコーダーモデルは抽象化された状態ベクトルを生成する（自然言語など元のものに復号しない）。このため、文書分類や感情分析などの選択問題、文章中の特定の単語やフレーズが何に相当するかを識別する固有表現抽出、文書間や文間の関係性判断（例えば整合的 / 矛盾 / 無関係）など、用途が限定されているという事情もあった[23]。

BERT で文章生成を行う場合は、デコーダー型の別のモデルを組み合わせるが、文章生成を目的とするならば、オリジナルの Transformer モデル通りにエンコーダー・デコーダーの一体型モデルで開発したほうが効率的である。この発想で登場したのが Facebook/Meta の BART（Bidirectional Auto-Regressive Transformer）であり、ほぼ同時期に Google も T5（Text-to-Text Transfer Transformer）を開発している（図中央の枝を参照）。

23：そうはいっても、技術的スタックは今も有用である。例えば、RAG では LLM の性能以上に情報データベースからの検索性能にサービスの質が依存するため、ベクトル化データベースに対する類似度検索でなく BM25 のような伝統的検索インデックス（TF-IDF の組み合わせ活用、詳細は打田他（2022）を参照）が重要となる。BM25 に BERT のアウトプットの一つである CLS（言語から符号化された状態ベクトルの一部で、意味情報などを集約している）を加味することで、検索能力を高性能化した BM42 が考案されるという動きが生じている（Vasnetsov 2024）。BM25/TF-IDF は 40 年前から利用され続けている技術であるが、そうした既存の技術との組み合わせなど LLM には様々な可能性が秘められている。

図表18　LLM進化の系統図



出所）　Clark et al.（2019）

Google は、BERT から T5 に移行し、その後は、PaLM/Gemini シリーズでデコーダーモデルに開発をシフトさせている。Facebook/Meta も、BART や BERT の発展形 RoBERTa から、2022 年以降は LLaMA シリーズや OPT のようにデコーダーモデルにシフトしている。Meta は、OpenAI や Google と異なりオープンソース化戦略をとっている点に特徴がある。

Google も Flan-T5 や Gemma など一部をオープンソース化している。Meta 社が OSS 提供している LLaMA シリーズをベースに開発を進めた LLM も多く登場しており、例えば、日本の直近の事例では ELYZA 社が Llama-3-ELYZA-JP-70B/8B を開発・公開している（70B はパラメータ数が 700 億であることを示している）[24]。

LLM を整理する場合、エンコーダー型かデコーダー型かというモデル構造上の視点のほか、オープンソースかクローズドか、大規模モデルか小規模モデルか[25]、学習方法のバリエーション、言語特化型か画像や音声にも対応したマルチモーダルか、学習データの性質（文章、プログラム言語、音声等）と量・質、API 対応の有無（商用は API 対応）、有償 / 無償、ファインチューニングの可否（次節参照）など、様々な視点がある。

24：最近のオープンソース LLM では、フランスの AI スタートアップ Mistral の Mixtral、UAE のテック企業 TII の Falcon、MosaicML の MPT、Hugging Face · BigScience プロジェクトの BLOOM、Databricks の DBRX、xAI（イーロン · マスクの X 関連企業）の Grok などが知られている。

25：LLM の大規模化が著しいため、大規模か小規模かは相対的な差に過ぎないが、ローカル LLM として PC やスマートフォンなどのデバイスでエッジコンピューティングが可能になるかという視点では、ハードウエア対比の絶対水準が意味を持つ。小規模 LLM のエッジコンピューティングは今後の発展方向として注目されている。最近の LLM は中小規模のものをシリーズで展開するケースが少なくない。大規模モデルで学習した内容を小規模モデルに転移させる技術として知識蒸留（Knowledge Distillation）と呼ばれる手法が発展しており、小規模モデルのラインナップ拡充に寄与している。

図表19　主要なLLM群

Type	Model Name	#Parameters	Release	Base Models	Open Source	#Tokens	Training dataset
Encoder-Only	BERT	110M, 340M	2018	-	✓	137B	BooksCorpus, English Wikipedia
	RoBERTa	355M	2019	-	✓	2.2T	BooksCorpus, English Wikipedia, CC-NEWS, STORIES (a subset of Common Crawl), Reddit
	ALBERT	12M, 18M, 60M, 235M	2019	-	✓	137B	BooksCorpus, English Wikipedia
	DeBERTa	-	2020	-	✓	-	BooksCorpus, English Wikipedia, STORIES, Reddit content
	XLNet	110M, 340M	2019	-	✓	32.89B	BooksCorpus, English Wikipedia, Giga5, Common Crawl, ClueWeb 2012-B
Decoder-only	GPT-1	120M	2018	-	✓	1.3B	BooksCorpus
	GPT-2	1.5B	2019	-	✓	10B	Reddit outbound
Encoder-Decoder	T5 (Base)	223M	2019	-	✓	156B	Common Crawl
	MT5 (Base)	300M	2020	-	✓	-	New Common Crawl-based dataset in 101 languages (m Common Crawl)
	BART (Base)	139M	2019	-	✓	-	Corrupting text
GPT Family	GPT-3	125M, 350M, 760M, 1.3B, 2.7B, 6.7B, 13B, 175B	2020		×	300B	Common Crawl (filtered), WebText2, Books1, Books2, Wikipedia
	CODEX	12B	2021	GPT	✓	-	Public GitHub software repositories
	WebGPT	760M, 13B, 175B	2021	GPT-3	×	-	ELI5
	GPT-4	1.76T	2023	-	×	13T	-
LLaMA Family	LLaMA1	7B, 13B, 33B, 65B	2023	-	✓	1T, 1.4T	Online sources
	LLaMA2	7B, 13B, 34B, 70B	2023	-	✓	2T	Online sources
	Alpaca	7B	2023	LLaMA1	✓	-	GPT-3.5
	Vicuna-13B	13B	2023	LLaMA1	✓	-	GPT-3.5
	Koala	13B	2023	LLaMA	✓	-	Dialogue data
	Mistral-7B	7.3B	2023		✓	-	-
	Code Llama	34	2023	LLaMA2	✓	500B	Publicly available code
	LongLLaMA	3B, 7B	2023	OpenLLaMA	✓	1T	-
	LLaMA-Pro-8B	8.3B	2024	LLaMA2-7B	✓	80B	Code and math corpora
	TinyLlama-1.1B	1.1B	2024	LLaMA1.1B	✓	3T	SlimPajama, Starcoderdata
PaLM Family	PaLM	8B, 62B, 540B	2022	-	×	780B	Web documents, books, Wikipedia, conversations, GitHub code
	U-PaLM	8B, 62B, 540B	2022	-	×	1.3B	Web documents, books, Wikipedia, conversations, GitHub code
	PaLM-2	340B	2023	-	✓	3.6T	Web documents, books, code, mathematics, conversational data
	Med-PaLM	540B	2022	PaLM	×	780B	HealthSearchQA, MedicationQA, LiveQA
	Med-PaLM 2	-	2023	PaLM 2	×	-	MedQA, MedMCQA, HealthSearchQA, LiveQA, MedicationQA
Other Popular LLMs	FLAN	137B	2021	LaMDA-PT	✓	-	Web documents, code, dialog data, Wikipedia
	Gopher	280B	2021	-	×	300B	MassiveText
	ERNIE 4.0	10B	2023	-	×	4TB	Chinese text
	Retro	7.5B	2021	-	×	600B	MassiveText
	LaMDA	137B	2022	-	×	168B	public dialog data and web documents
	ChinChilla	70B	2022	-	×	1.4T	MassiveText
	Galactia-120B	120B	2022	-		450B	
	CodeGen	16.1B	2022	-	✓	-	THE PILE, BIGQUERY, BIGPYTHON
	BLOOM	176B	2022	-	✓	366B	ROOTS
	Zephyr	7.24B	2023	Mistral-7B	✓	800B	Synthetic data
	Grok-0	33B	2023	-	×	-	Online source
	ORCA-2	13B	2023	LLaMA2	-	2001B	-
	StartCoder	15.5B	2023	-	✓	35B	GitHub
	MPT	7B	2023	-	✓	1T	RedPajama, m Common Crawl, S2ORC, Common Crawl
	Mixtral-8x7B	46.7B	2023	-	✓	-	Instruction dataset
	Falcon 180B	180B	2023	-	✓	3.5T	RefinedWeb
	Gemini	1.8B, 3.25B	2023		✓	-	Web documents, books, and code, image data, audio data, video data
	DeepSeek-Coder	1.3B, 6.7B, 33B	2024	-	✓	2T	GitHub's Markdown and StackExchange
	DocLLM	1B,7B	2024	-	×	2T	IIT-CDIP Test Collection 1.0, DocBank

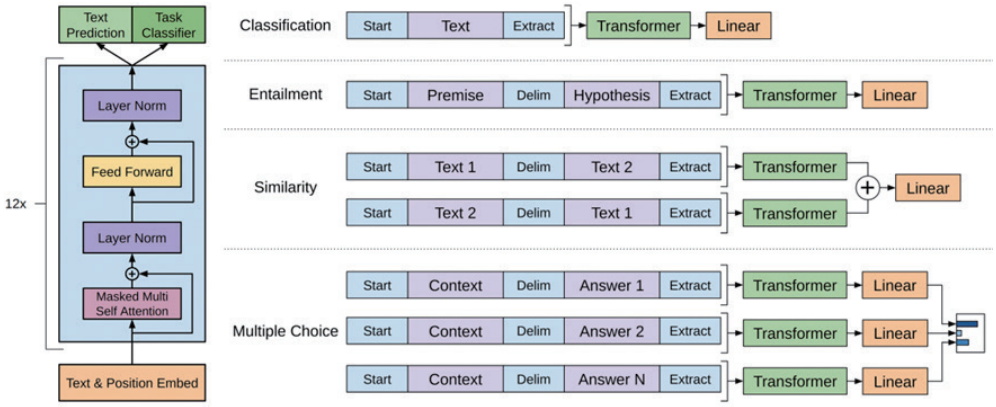
出所）Minaee et al.（2024）

図表 20 は GPT-1 のデコーダーモデルを示したものである[26]。Attention 機構の内部は同じであり、12 階層となっている。デコーダーの Attention 機構では未来情報にマスクがかかっているため、既出である単語について過去方向にしか Attention を向けることができない分、学習能力は低下するが、BERT と異なり文章などへの復号が可能という利点がある。単語の分散表現に BPE（Byte Pair Encoding）と呼ばれる手法を用いており、単語を更にサブワードに分割してトークン化し、これを分散表現の対象とするものである（例えば、sub/word、富士 / 山、もしくは富 / 士 / 山）[27]。

26：図表の右側では、インプットとして 2 文を接続させ、セパレーターで挟み込んで使う方法が示されている。これは、同時期に開発公表された BERT も同様な作りとなっている。文書分類なら 1 文のみ、命題と仮説からの含意は 2 文の組み合わせ、類似性評価なら 2 系統構成というように使い分けが示されている。

27：日本語の文章を分散表現にかけると、富 / 士 / 山のように、ほぼ一文字が 1 トークンになるまで分解される。英語のように単語間でスペースが存在する言語に比べ、切れ目がない言語は一般に非常に細かくトークン化される傾向がある。一文字が複数のトークンで表現されることもある。

図表20　GPT-1のデコーダーモデル



出所）Radford et al.（2018）

LLM 開発においては性能評価が必須であり、評価手法についても研究開発が進んでいる。図表 21 は、AI/ML 開発者にモデル開発プラットホームを提供する Weights & Biases の日本法人が提供している日本語を対象としたリーダーボードである。llm-jp-eval（言語理解）と Japanese MT-bench（言語生成）で評価されていた「汎用的言語性能」に加え、「アラインメント」という新たな評価軸が導入され、総合評価上位順にリストアップされている。Weights and Biases Japan（2024）はこれらの評価軸について解説しており、汎用的言語能力が翻訳や推論、検索、知識・質問応答、意味解析など様々な視点からの能力評価の合成値であることを解説している。アラインメントは、制御性、倫理・道徳、毒性、バイアス、真実性、堅牢性などといった公正性（フェアネス）などに係る視点からの評価がなされている。

図表 21 は、同社のサイトに掲示されているリーダーボードであり、ユーザが表示形式を操作可能な UI（ユーザインターフェイス）が採用されている。横軸に様々な評価基準が並んでおり、汎用的言語性能でリストアップすると、上位には Anthropic 社の Claude3.5 sonnet や OpenAI 社の GPT-4o、Gemini 1.5 Pro などの 2024 年 7 月初時点の最新の商用 LLM がランクインしていることが見て取れる。図表 22 は同社のブログ（note）に掲載されたものであり、上記 2 軸で評価された LLM が x-y プロットされている。グローバル企業の商用 LLM（ほとんどは API 提供されており外部から使用）が両軸と

もに上位となっていることがわかる。ランキングは日進月歩で変化しており、同社 website で時系列プロットしてみると性能向上が日々続いていることがわかる。

図表21　Weights & Biases JapanのNejumi LLMリーダーボード3

GLP : General Language Processing (汎用的言語性能)

ALT : Alignment (アライメント)

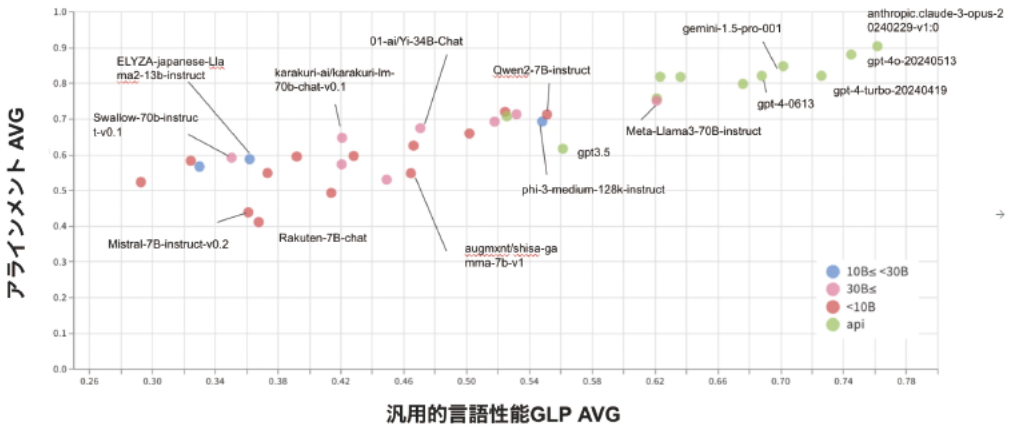
Total AVG = (Avg. GLP + Avg. ALT)/2

runs.summary["leaderboard_table"]

	model_name	model_size_cz	汎用的言語性能(GLP)_AV	アラインメン	TOTAL_AVG	GLP_表現	GLP_翻訳	GLP_情報検索	GLI
12	anthropic.claude-3-5-sonnet-20240620-	api	0.7618	0.9027	0.8322	0.8733	0.8684	0.8061	
6	gpt-4o-2024-05-13	api	0.7451	0.8796	0.8123	0.895	0.8693	0.8227	
35	gpt-4-turbo-2024-04-09	api	0.7019	0.8467	0.7743	0.8917	0.8607	0.813	
13	anthropic.claude-3-opus-20240229-v1:0	api	0.7262	0.82	0.7731	0.885	0.872	0.7899	
15	gemini-1.5-pro-001	api	0.6882	0.8199	0.754	0.88	0.8534	0.8019	
36	gpt-4-0613	api	0.6762	0.7973	0.7367	0.8133	0.8587	0.8926	
10	gemini-1.5-flash-001	api	0.6365	0.817	0.7268	0.8667	0.8452	0.7472	
9	anthropic.claude-3-sonnet-20240229-	api	0.6235	0.8176	0.7205	0.8567	0.8313	0.748	
14	anthropic.claude-3-haiku-20240307-v1:0	api	0.6214	0.7564	0.6889	0.8183	0.8045	0.7991	
31	meta-llama/Meta-Llama-3-70B-Instruct	30B≤	0.6213	0.7499	0.6856	0.7533	0.8516	0.8753	
1	cyberagent/calm3-22b-chat	10B≤ <30B	0.625	0.7164	0.6707	0.8517	0.8431	0.8825	

出所）Weights & Biases Japan Website 2024年7月5日時点

図表22　日本語LLMリーダーボード



出所）Weights & Biases Japan（2024） https://note.com/wandb_jp/n/nd4e54c2020ce

4．LLM発展過程での発見

4.1 3つのスケール測と規模拡大競争

学習や推論において並列処理が可能となった Transformer 型のモデルでは、モデルの大規模化が進展した。Attention 機構により文章上、遠く離れた箇所どうしの関係性が補足できるようになったため、長文をまとめて処理する利点を活用したいという誘因も大規模化を促した。インプット数の増大は、DNN の上部の層が深いほどパラメータ数を乗数的に増加させる。ブロックの繰り返しによりパフォーマンスを引き上げられる点もディープ化を促した。また、そもそも Attention 機構の QKV の仕組みは大量にパラメータを使用するものであった。一方で、並列処理による計算負荷の軽減が、大規模化に伴う計算制約を緩和する効果もあった。

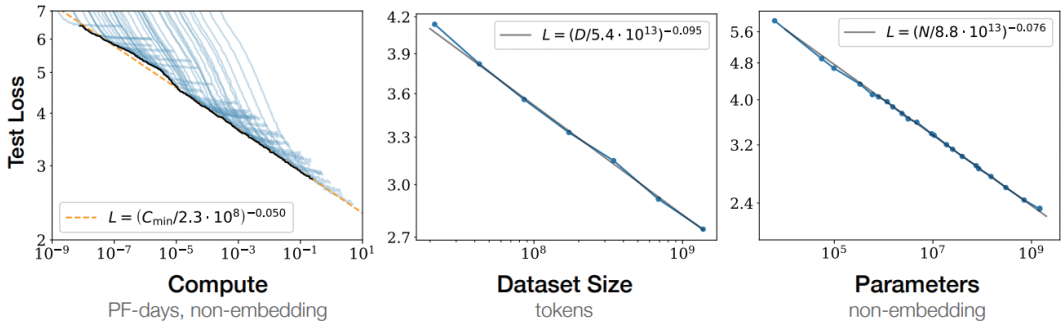
このようなモデル発展上の要請のほか、GPU などハードウエアの進歩、クラウドコンピューティングの普及といった計算資源の拡充や、分散コンピューティングの技術進歩による分散学習（並列処理の利点）、学習用データセットの整備（Common Crawl などを活用したインターネット上の情報収集など）もモデルの大規模化を促した。

Kaplan et al.（2020）は、OpenAI の研究者 10 名による規模とパフォーマンスの関係に関する調査であり、計算資源、学習データ、モデルサイズ（パラメータ数）の 3 要素を大規模化させた場合、どの程度のパフォーマンス改善が図れるかを各要素について検証している。同論文は、図表 23 に示された 3 つのスケール則を発見している。右図は、モデルを大規模にしてパラメータ数を増やすほど、べき乗則に従ってモデルのパフォーマンスが高まることを示している（縦軸の損失関数が線形に減少[28]）。中央図は学習に用いるデータセットの規模についてもべき乗則のスケール則が成立していることを示している。左図の計算資源を対象とした検証図には複数の右下がり線が描かれている。図中の右側の線は、パラメータ数が多いモデルに計算資源をより多く投入した場合のパフォーマンス改善を示している。線形に改善していくが一定値までくると限界に直面し、右横方向に推移している（正確には緩やかな右下がりに変化）。左側の中小規模モデルは、一定のパフォーマンスを達成するには大規模モデルより計算資源を必要としない（y 軸の一定値で比較）が、改善の限界に直面するのも早く、それ以上の改善を望むならモデルを大きくするしかないことを示している。この限界線を様々な規模のモデルについて結ぶと右下がりの直線（黒太線）が現れており、計算資源についてもべき乗則が成立していることがわかる。

この発見は、LLM の開発方針に大きな影響を与えた。大規模化すれば確実に性能が改善することが、べき乗則が成立し続ける間は保証されている。図表 24・25 は GPT シリーズのパラメータ数の拡大を示している。GPT-2 から GPT-3 に向けてパラメータ数が 100 倍以上に引き上げられ、学習データもより巨大なものを用いるようになり、学習に巨大な計算資源を振り向けられるようになった。この時期に、LLM の開発における規模拡大競争が明確なものとなった。

28：べき乗則は、べき分布（power law distribution）に従って「ある値が他の値に対してどのようにスケーリングするか」を示すものである。べき分布は $P(x)=cx^{-\alpha}$（α はべき指数）と表現され、これを対数表示すると、$\log(P(x))=\log(C)-\alpha\log(x)$ となる。同式は、x が 1%増加すると $P(x)$ が α %減少することを意味している。図表 23 は、表示された領域全域で x の規模にかかわらずべき乗則が維持されていることを示している。なお、べき乗則は、家計の所得や資産分布、市町村の人口規模の分布、書籍の販売部数の分布、資産価格変動率の分布といった社会現象だけでなく、自然現象においてもしばしば観察され、フラクタル理論とも関連性が高い概念である。例えば、与信ポートフォリオの集中リスク計測においても、企業規模 / 与信規模の分布は重要な要素となる。金融市場の価格変動にもべき乗則やフラクタル特性が観察されている。

図表23　3つのスケール測



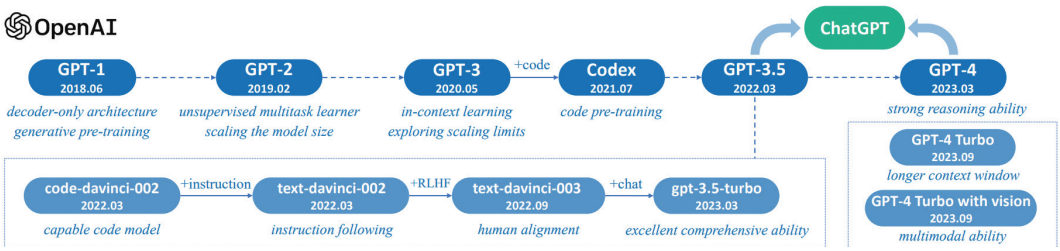
出所）Kaplan et.al.（2020）

図表24　GPTシリーズの規模拡大

モデル	パラメータ数	公開年月	主な特徴、開発論文
GPT-1	1.17 億	2018/6	Transformer のデコード型モデルとして言語生成を行う事前学習（pre-training）型の初の大規模言語モデル。Radford et al.（2018）
GPT-2	15 億	2019/2	Attention 機構を 48 層としモデル規模を拡大させたほか、大規模データセットで学習することにより、一貫性がある文章が生成されやすくなるなど能力が向上。Radford et al.（2019）
GPT-3	1,750 億	2020/5	スケール則の発見を活かすべくパラメータ数を一気に約 100 倍増させ、言語生成能力が大きく向上。Common Crawl という大規模データセット（後述）を学習に利用している。In-context learning が可能となった。Brown et al.（2020）
GPT-3.5	非公開	2022/3	改良されたトレーニング方法とデータセットにより、性能と応用範囲が更に向上し、ゼロショットでの回答精度やアラインメントの質が高まったことから対話サービス ChatGPT のバックエンドで利用されるようになった（図表 25 参照）。
GPT-4	非公開	2023/3	マルチモーダルに対応。精度と安全性を向上させるため、強化学習と人間によるフィードバックを利用している。Open AI et al.（2023）

出所）各開発論文より

図表25　GPTシリーズの発展とサブライン（下段）

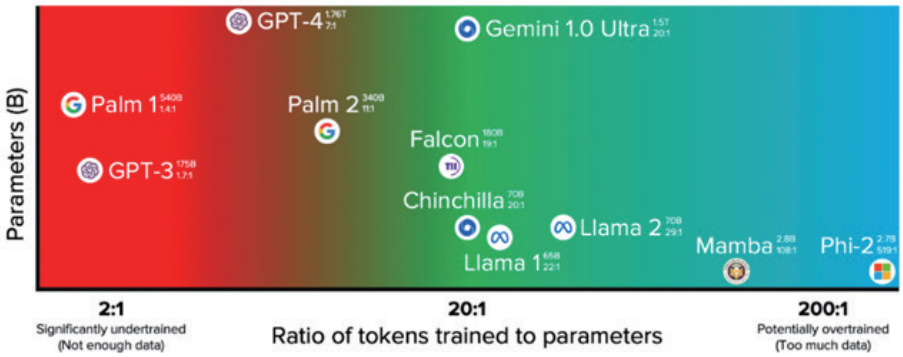


出所）Zhao et al.（2023）

前述のスケール則を前提とすると、3 つの要素のうち何を拡張すると効率よく性能向上が図れるかという問題に直面する。現状を前提に改善するときは、スケール則に従ってパラメータと学習データセットの増加率を同じにすればよい。しかし、現在の状態がベストなバランスである保証はない。DeepMind の研究者たちは、Hoffmann et al.（2022）において、計算資源を一定とすれば、最良のモデルは必ずしも大きいモデルではなく、比較的小規模なモデルを比較的大量のデータで学習したモデルとなることを示し、当時加速しつつあったモデルの巨大化開発方針が適切でない可能性を指摘した。同論文は、トークン（単語）数とパラメータ数の比率は 20：1 が望ましいとしている。この目安の算出には同社が開発した LLM の Chinchilla を検証に用いたため、チンチラのスケーリング則と呼ばれるようになった。その後、Sardana and Frankle（2023）は、チンチラ則は学習時の計算資源制約を前提としたものであり、実際に使用する場合の推論コスト（inference cost）は考慮していないという指摘を行い、更にモデル規模を小さくしたほうが望ましい（190：1）という検証結果を得ている。実用化の際には、ランニングコストが重要になるため、推論コストを検討対象に入れることは重要である。また、生成速度というレスポンス時間の面でも最適規模の選定は実用化の際の重要な要素となる。

ただし、その後も、先端的モデルの大規模化は継続しており、GPT-4 は非公表ながら数千億から 1 兆を上回っているという推測がなされている（図表 26 では 1.76 兆という数字が示されている）。他の主要モデルも大規模化を図っており、図表 27 のリーダーボード（日本語評価）の上位に出てくるようなモデルは数百億から数千億のオーダーとなっている。図表 26 に示した主要 LLM のトークン数とパラメータ数のバランスもチンチラ則からは様々に乖離している。例えば、大幅な大規模化を目指した GPT-3 は 20：1 ではなく 2：1 でモデル規模を優先させている。

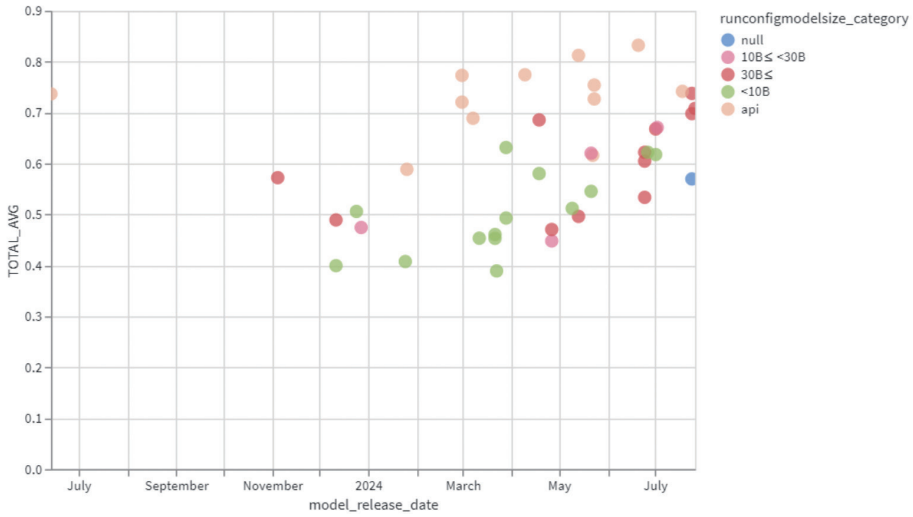
図表26　チンチラ則との乖離



出所）Thompson（2024）

一方で、コスト対比でのパフォーマンスを優先させるため、ハイエンドモデルを知識蒸留などにより小型化したモデルも多く登場している。オープンソースの LLM を専門分野知識特化型や特定言語特化型に拡張開発するため、別学習セットを用いて追加学習させる研究も盛んに行われている。そこでも開発・学習コストの面で、数十～数百億程度の中小型モデルが採用されている。日本語を強化したモデルとしてリーダーボードに登場している開発例をみても、サイズを抑制する代わりに学習データセットとチューニングやアラインメントで性能を上げる方針のものが多い。前述したリーダーボードのサイトで確認すると ELYZA/Llama-3-ELYZA-JP-8B や、Cyberagent/calm3-22b-chat、KARAKURI-AI/karakuri-lm-8x7b-chat-v0.1、TokyoTech-LLM/Llama-3-Swallow-8B-Instruct-v0.1 などである。これらの多くは、Llama-3 や Calm3 などオープンソースの最新 LLM をベースに追加学習が行われている。一方で、NEC や Preferred Networks、ABEJA などのようにスクラッチで LLM を開発している先もあり、これらについてもグローバルトップリーダーの最新 LLM と比較すると比較的規模が小さいモデルを用いて日本語対応性能を引き上げようとするものが多い。

図表27　Weights & Biases JapanのNejumi LLMリーダーボード3

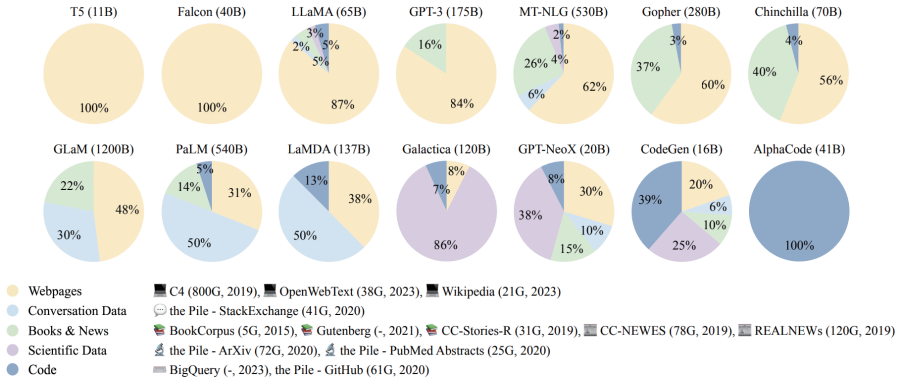


出所）Weights & Biases Japan, Nejumi LLMリーダーボード, 2024年7月31日取得

学習セットの大規模化は、量と対象の面において拡大している。まず、インターネット上にある情報をクローリングして集めてくる手法の高度化があげられる。Common Crawl は、ペタバイト規模のデータ量を含むオープンソースのウェブクロールデータベースであり、LLM の学習に広く利用されている。サブセットも整備されており、例えば、C4、CC-Stories、CC-News、RealNews などがある。C4（Colossal Clean Crawled Corpus）のなかにおいても、en（806G）、en.noclean（6T）、realnewslike（36G）、webtextlike（17G）、および multilingual（38T）の５つのバリエーションがある（カッコ内はバイト数表示＜ギガバイト、テラバイト＞、トークン化を行った後はトークン数で表記されることが多いが、コーパス段階ではバイト評

価が一般的）。このほか、各国語の Wikipedia も生コーパスとして利用されている。Reddit は、Facebook とほぼ同時期に誕生した米国発祥の SNS プラットホームで、匿名で書き込まれ米国版 2 ちゃんねるとも呼ばれているが、文章の質が全般に保たれているため、高品質のデータセットを作るための生コーパスの元になっている。著作権が切れた書籍も学習用のデータベースとして利用されており、BookCorpus や Project Gutenberg が知られている。コンピュータプログラムも学習対象であり、GitHub のようなコードレポジトリや、StackOverflow のようなコードに関する Q&A プラットホームも学習用データとして整備・活用されている。図表 28 は 14 の LLM について、学習データセットの構成比とサイズ（LLM に付されている数値はトークン単位、凡例のデータセットはビット単位）を示している。

図表28　学習データセットの巨大化と多様化

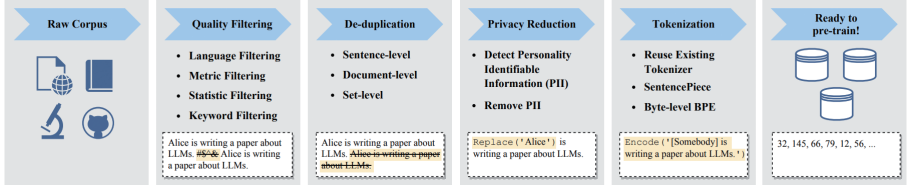


出所）Zhao et al.（2023）

規模以外にも学習データセットの質が LLM の性能に影響するため、図表 29 で示したような前処理が行われている[29]。生コーパスからの品質フィルタリング、重複の除去、プライバシー情報の削除（これはアラインメント目的）、html のタグ除去（一部には文脈上の意味があるものも存在）、分散表現を行う前のトークン化（文を単語や単語以下のサブワードに分割する）などが行われたうえで、分散表現技術によりベクトル化される。こうした過程で生コーパスの規模（ビット計測）は大きく縮約される。例えば、Penedo et al.（2023）を参照。また、学習データの質を高めるほうが規模を大きくするより遥かに効果的であるという実証研究もなされている（Gunasekar et al. 2023）。

29：日本語大規模ウェブコーパスについて同作業を行ったものとして岡崎他 (2024) がある。

図表29　学習データの生成



出所）Zhao et al.（2023）

4.2 様々な転移学習の発見：事前学習とファインチューニング

LLM の発展を理解するためには、転移学習が自然言語処理の発展に果たした役割を知ることが重要となる。転移学習とは AI や機械学習の発展において生じた概念であり、ある領域（ソースとなるドメイン）のタスクにおいて獲得された知識を別の領域（目標ドメイン）のタスクに転用することを指す。松井・熊谷（2024）が転移学習について包括的な解説を行っている。生成 AI/LLM における転移学習は、発展段階の様々なポイントで異なった形態で登場してきた。以下で順に解説する。

4.2.1 分散表現

最初の代表事例は、word2vec のような単語（トークン）の分散表現である。あるコーパスから学習によって得られた分散表現は、一定の汎用性をもって別の領域の自然言語処理にも適用できる。これは転移学習の一形態に相当する。前出の ELMo は文脈に依存して単語を分散表現する手法であり、こちらも文脈に関する獲得情報が転移学習可能であることを示している。LLM 開発において学習データセットは英語がほとんどであり他国語は少ししか含まれていない場合であっても、学習済みの LLM が他国語への対応能力を一定程度有することが知られている。言語の成り立ちに普遍性があるならば、主に英語から抽出された言語の特徴が他国語の自然言語処理にも一部転移可能であることを示唆している。

4.2.2 事前学習とファインチューニング

GPT-1 と BERT は、事前学習（Pre-training）とファインチューニング（Fine-tuning）というコンセプトを LLM に導入した[30]。自然言語理解や生成において転移学習が可能であれば、ある LLM が大きな学習データセットにおいて獲得した自然言語処理能力を別の LLM に転移することができる。この前半部が事前学習であり、その転移された能力を更に改善する行為をファインチューニングと呼ぶ。一般に、ファインチューニングにおいてはモデルのパラメータ群が更新されるため、追加的な学習データセットが比較的小規模であったとしても大きな計算資源を要することが多い。チューニングも容易ではなく、高精度・高性能に作りこまれた汎用 LLM の性能を逆に落としかねないリスクもある。機械学習、とりわけ NN では、新しい情報を学習する過程において以前に学習した情報が急激に失われる現象が生じることが知られている。これは破滅的忘却（Catastrophic Forgetting）と呼ばれ、LLM の事前学習プロセスやファインチューニングにおいても発生している[31]。こうした現象を回避し、かつ膨大な計算資源の消費を回避するために、より効率的な追加学習方法として PEFT（Parameter Efficient Fine-Tuning）と呼ばれる技術が開発されており後述する。

4.2.3 教師ありファインチューニング

より簡便な方法として、教師ありファインチューニング（Supervised Fine-Tuning）がある。事前学習されたモデルを特定のタスクやドメインに

30：Radford et al.(2018)、Devlin, Chang, Lee and Toutanova(2018) を参照。

31：2022 年にリリースされた GPT-3.5 ではユーザによる追加学習が可能となった。筆者は専門領域分野の日本語 Q&A100 問を用いて小規模な Supervised Fine-Tuning を GPT-3.5 で行ったことがある。専門知識の反映はある程度の精度を持って行えたが、回答文章の日本語品質が明らかに低下しており、「破滅的」忘却ほどではないが汎用 LLM の機能が一部劣化することを経験した。追加学習にかかるコスト（計算負荷に応じてチャージされる利用料金）も推論に比べると高い。

適応させるために、正解ラベル付きデータ（教師付きデータ）を追加的に用いてモデルの学習を進める手法である。同手法では、追加学習させたいタスク（転移学習でいう目標ドメイン）に関連するラベル付きデータを作成する手間が必要となる。

4.2.4 RLHF：Reinforcement Learning from Human Feedback

強化学習（Reinforcement Learning）は、エージェントが環境との相互作用を通じて報酬を最大化する行動方針を学習する方法である。強化学習は LLM でも採用されており、RLHF（Reinforcement Learning from Human Feedback）が典型例である。まず、人間のアノテーター（何らかのラベルを付ける人）が LLM の出力を評価し、その評価に基づいて報酬モデルを訓練する。次に、この報酬モデルを用いて、LLM が新たに生成する出力を評価し、強化学習アルゴリズム（例えば、Proximal Policy Optimization; PPO）を通じて LLM をチューニングする。こうした手法により、LLM が人間（評価者）の好みに合致する応答を生成する能力を高めることができる。人間のフィードバックを強化学習に介在させるため、Human Feedback という名前が付いている。LLM は事前学習で得た情報をもとに新しい状況やタスクに評価関数を通じて適応していくため、RLHF は転移学習の一形態とみなされる[32]。GPT-3.5 の開発においては、図表 25 の下段にあるシリーズ（InstructGPT）で RLHF が採用され、GPT-3 からの能力改善を果たしており、これが ChatGPT のチャット能力を高めた。Anthropic の Claude や DeepMind の Gopher においても RLHF が活用されている。

32：松井・熊谷（2024）は 13 章で強化学習一般における転移学習を解説している。

4.2.5 ICL：In-Context Learning

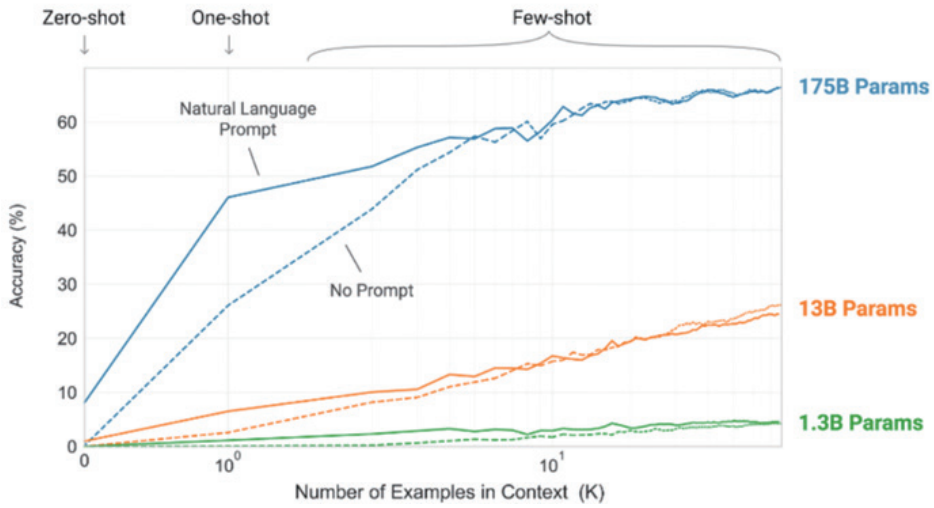
ファインチューニングや追加学習では、LLM のパラメータ再設定が行われるが、LLM のパラメータを不変としつつ転移学習を活用するコスト効率の高い手法が考案された。現在のプロンプトエンジニアリングに繋がる In-Context Learning（ICL）である。ICL は GPT-3 の開発において発見された。具体的には、LLM の推論時に、インプット指示（プロンプト）の補足情報として望ましい回答事例や回答様式を示すことで、優れた回答を引き出す手法である。これにより、事前学習された知識を新しい状況やタスクに適用することが可能となる。

ICL は、文脈に沿った指示を新しいデータとして与え、LLM が特定のタスクを遂行する能力を高める手法であり、転移学習が新しいタスクに対してモデルを適応させるメカニズムと類似している。また、ICL は、モデルが事前に学習した一般的な知識を活用し、特定のタスクに対して柔軟に適応することを可能にする。例えば、少量のデータで LLM に高性能を発揮させ、迅速に新しいタスクに適応させることができるという利点がある（追加学習が不要）。このようにプロンプトに含まれる少ない情報だけで LLM がタスクを遂行できることは、転移学習の持つ適応能力を反映したものである。

ICL では、事例を一つだけ与えるケースを One-shot learning（あるいは One-shot プロンプト、以下同じ）と呼び、数個の場合を Few-shot learning

と呼ぶ。ここから派生して、回答事例なしでプロンプトを与えるのを Zero-shot learning と呼ぶようになった。プロンプトエンジニアリングでは、ICL だけではなく、LLM に回答者としての立場（役割）を与えたり、特定の口調やスタイルを指示したりすることも含まれる。関連する背景情報や文脈をプロンプトに含めることで、より正確で関連性の高い応答を生成できるようになる。例えば、過去の会話履歴や特定の知識ドメインに関する情報をプロンプトに追加する事例があげられる。図表 30 は、GPT-3 において発見された ICL の性能改善を数値検証した研究内容を示している。モデル規模ごとに ICL の効果が表れていることや（小規模モデルでも右上がりとなっており、規模差なりの改善を示している）、大規模モデルでは One-shot learning だけで大きな性能改善効果があることが観察されている。

図表30　GPT-3でのOne-shot/Few-shotでの性能改善



出所）Brown et al.（2023）

4.2.6 マルチモーダル

転移学習は LLM のマルチモーダル化でも活用されている。多くのマルチモーダルモデルは、既存の単一モーダル（例えば文章のみ）で学習されたモデルをもとにしており、その知識を利用して他のモーダル（例えば、画像や音声）を理解・生成する能力を追加する。例えば、OpenAI の CLIP（Contrastive Language-Image Pre-training）は、大規模なテキストデータセットで事前学習された言語モデルと、画像データセットで事前学習された視覚モデルを組み合わせており、テキストと画像のペアを用いて学習することで両者の関係を理解する能力を獲得している。

転移学習を用いることで、テキストと画像など異なるモーダル間の知識を統合し、それぞれのモーダルが持つタスク実行能力を相互に補完することができる。例えば、DALL-E や Flamingo のような生成 AI モデルは、テキストから画像を生成する、あるいは画像を理解してテキスト表現する能力を有している。

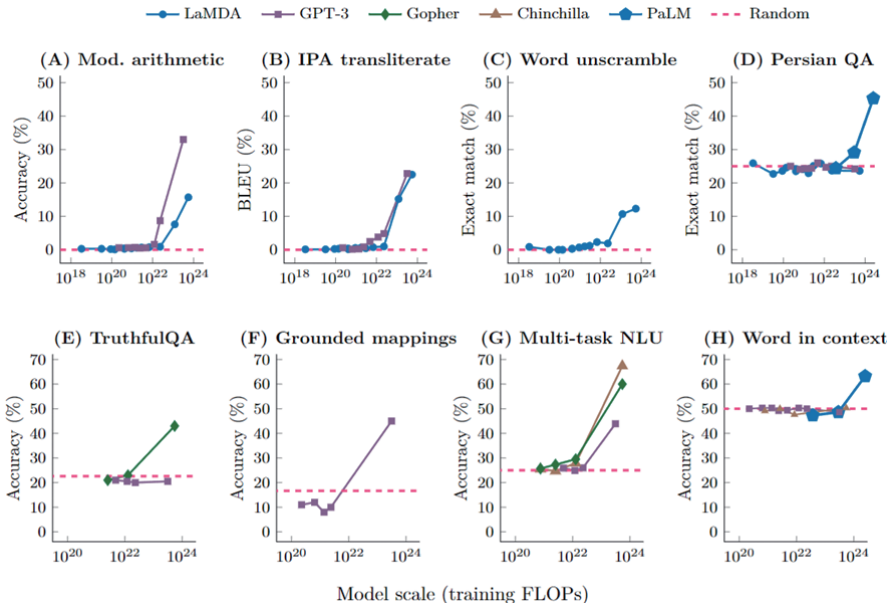
このように元々は文書の生成 AI であった LLM に対して、画像や音声の生成 AI の知識やスキルを転移統合することが行われており、LLM は文章生成 AI を超えて高い汎用性を持つ生成 AI 化してきている。このため、生成 AI と LLM の区分は曖昧化しつつある。

4.3 創発現象

LLM の大規模化に伴って発見された点に、創発現象（emergent ability）がある。これは、LLM の文脈では、モデルが小規模な間は対応することができなかった問題やタスクが、規模がある水準を超えると突然解けるようになる（能力が出現する）ことを指す。前述の In-context learning がその一例である。Wei et al.（2022）は他の様々な創発事象を取り上げている。図表 31（A）では、3 桁の加算と減算、2 桁の乗算をテストする算術ベンチマークの検証結果が示されている。横軸に示されたモデル規模がある閾値を超えると突然 LLM が計算能力を獲得することがわかる。発音記号から単語をスペルアウトする問題（B）、単語内でアルファベットの順番をランダムに入れ替えたものから元の単語を復元する問題（C）、ペルシャ語の質問応答（D）、質問に対する真偽の判定（E）、方位などの概念を扱う能力（F）、数学、歴史、法律など多様なトピックをカバーするテスト（G）でも、同様な創発現象が観察されている。

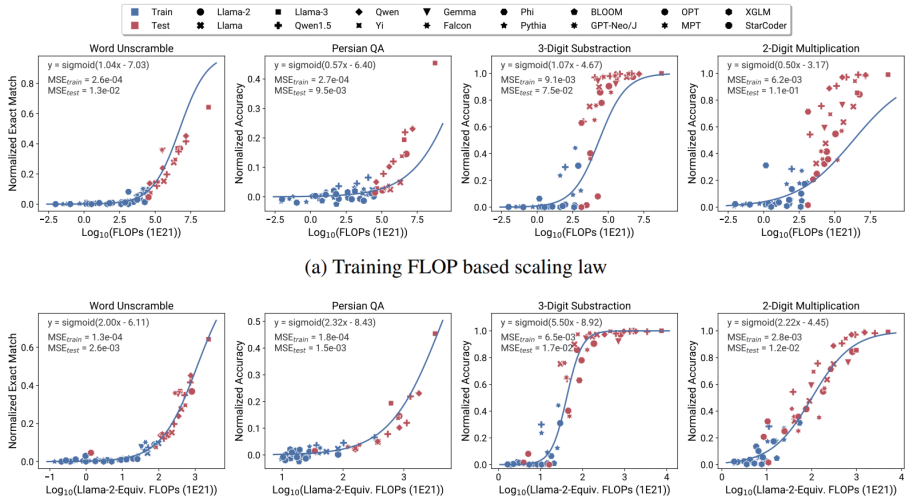
LLM の様々なタスクに関する能力の非連続的な変化は、モデルやベンチマークテストの開発に不確実性をもたらし、開発方針や評価を困難なものとする（べき乗則の逆ケース）。これらの能力が本当に非連続的であるのか、使用される評価指標の癖によるものなのか、あるいは検証の際の解像度が不足しているせいなのか、議論や検証がなされている。Ruan, Maddison and Hashimoto（2024）は、解像度を上げて詳細に見れば突然変化しているのではなく徐々に能力は上がっていることを指摘している（図表 32）。このほか、Schaeffer, Miranda and Koyejo（2024）は、評価指標の設計や選択によって突然能力が上昇しているように見えているのであり、指標を変えると連続的な変化になりうることを指摘している。ただし、今できないことが、どこまで大規模化したらできるようになるかという不確実性の問題は引き続き解決されていない。

図表31　創発現象の発生



出所）Wei et al.（2022）

図表32　創発発生地点の高精度計測



出所）Ruan, Maddison and Hashimoto（2024）

5．RAGとFine-tuning、AIエージェント

5.1 汎用LLMの限界と３つの対応法

４節でみてきたように、LLM は高性能化を続けている。しかし、企業内での利用や顧客情報を活用したサービス提供に生成 AI/LLM を利用する場合、内部情報や秘匿すべき情報をどのように活用するかという問題に直面する。そうした情報は「汎用」LLM の学習情報セットには含まれていない（含めてもいけない）。また、汎用 LLM の学習には膨大な計算資源が必要となり、ユー

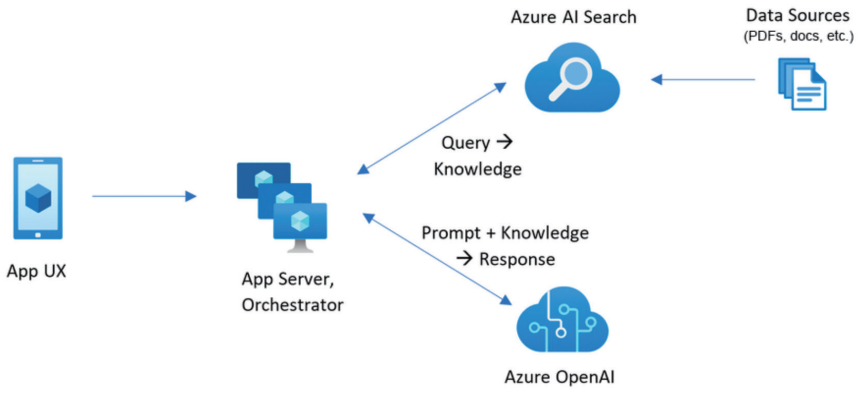
ザ企業にとって追加学習を行うことは容易ではない。開発企業ですら学習内容の頻繁なアップデイトは行われておらず、通常、カットオフデイトが設定され、その時点までで得られた情報に基づく学習が行われている。更には身の回りの細かい粒度の情報や頻繁に更新される情報を LLM に学習させるのは著しく非効率的であり、そのようなことは行われていない。それゆえ、今日の東京の天気や社内規定や、近所にあるお勧めレストランを聞いても汎用 LLM は正しく回答できず、ハルシネーションを起こすか、情報がないと回答するだけである。

こうした問題に対して、①非公開情報 / 秘匿情報を LLM の外部にデータベースとして持たせる、②コスト効率的なファインチューニングによりこれらの情報を追加学習させる、③一般公開情報を Web サイトで検索したり、特定情報を提供するサービスサイトから取得してくる、というアプローチが考案された。5 節ではこれらを順に説明する。

5.2 RAG

情報を外部データベースに保有させるアプローチとして、RAG（Retrieval Augmented Generation、検索拡張生成）という手法が Lewis et al.（2020）によって提唱された。①利用対象となりうるような情報をデータベースとして LLM の外部に構築し、② LLM への問いかけに対して関連性が最も高い情報をデータベースから抽出する。③これを LLM への問いかけ情報とセットにして、その情報の範囲内で回答するようなプロンプト指示と合わせて LLM に送る、というシステムである。図表 33 はその構成の一例である。これにより、汎用 LLM が一般情報に基づいた回答を返してしまいハルシネーションや不適切な回答が生成される可能性を抑制することができる。同時に、非公開情報 / 秘匿情報を活用したサービスを情報にアクセスする権利がある適格者だけに提供することが、システム上の工夫によって可能となる。RAG においては、LLM は、知識のストレージとアウトプット機能というより、与えられた情報に基づいてプロンプトの要求に従って回答を作成する自然言語生成機、すなわち言葉を操れる便利な UI として機能している。

図表33　RAGのシステムアーキテクチャの事例



出所）Azure website

RAG システムの構築においては LLM 以外の IT 知識が必要となる。LLM の性能以上に、情報データベースを的確に構築し、そこから回答生成に必要な情報を不足なく検索・収集してくる検索システムの性能が鍵となる。情報検索においては、ベクトル化データベースを構築し、問い合わせ情報を分散表現したうえで、類似度が高い（内積が大きい）ものをベクトル化データベースから選択する方法が LLM の自然な延長線上で発展してきた。しかし、ベクトル化技術 / 分散表現技術にこだわる必要は必ずしもなく、伝統的な検索技術も応用可能である。データベースをインデックス化し、検索用の指標、例えば TF-IDF（特定の文書における特定の単語の重要性を評価するための指標）やその改良版の BM25 を用いて、適切な情報をインデックス検索する手法である[33]。Microsoft 社のクラウドサービス Azure では、検索サービスとして Azure AI Search（図表 33）を提供しているが、ベクトル検索とインデックス検索に加えてセマンティック検索の 3 つをハイブリッドで用いて検索精度を改善させている。検索精度の向上は RAG のパフォーマンスに直結する。

33：RAG の隆盛により以前より発展してきた検索技術に再び注目が集まっている。検索技術の詳細は、例えば打田他（2022）に最新技術まで含めた解説がある。

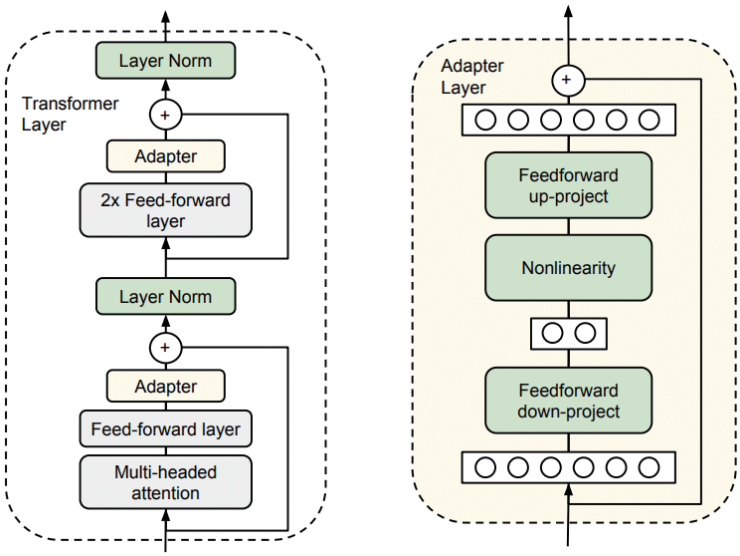
2023 年には、企業が顧客向け、社内向けに生成 AI サービスを構築する動きが加速したが、RAG は、2024 年上期においても主要な開発手法となっている。これは後述する PEFT が、効率化されているとはいえ Fine-tuning で LLM のパラメータを再推計するという大がかりかつ精度を引き上げるのが容易ではない手法であるのに対し、RAG においては情報データベースの構築と検索精度が確保できれば、LLM 自体をエンジニアリングする必要がないからである。そのため、RAG 構築を支援する様々なサービスが提供されており、6 節では簡単な事例を用いて複数の RAG 構築例を紹介する。

5.3 PEFT

2018 年に GPT-1/BERT が事前学習と Fine-tuning の組み合わせを提唱したが、その後の LLM の大規模化加速で Fine-tuning のコストがより大きなものとなり、開発企業でもない限り現実的な選択肢ではなくなった。そこで、事前学習で獲得されたパラメータは固定したうえで、モデルの一部追加と当該部分のパラメータを調整することで Fine-tuning を実行しようというアプローチが登場した。これらの技法は PEFT（Parameter Efficient Fine-Tuning）と呼ばれる。

Houlsby et al.（2019）はアダプター層（adapter layers）と呼ばれる追加モジュールを既存の事前学習モデルに挿入し、アダプター層のみをファインチューニングする手法を提案した。実際に BERT と複数の学習データセットで改善度合いを検証し、同アプローチが有効なことを確認している。アダプター層はフィードフォワード型の NN に類似した構造となっており、図表 34（右図）のように中間層の次元を圧縮することでパラメータ効率性を高めている。

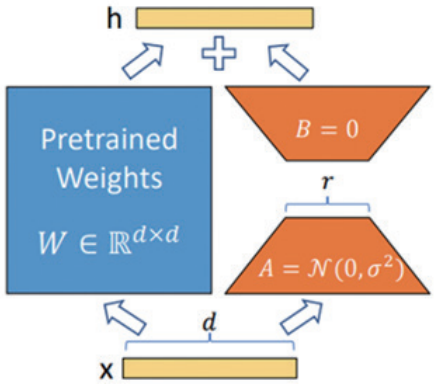
図表34　アダプター層の追加と内部構造の例



出所）Houlsby et al.（2019）

Hu et al.（2021）は、Transformer モデルの Attention 機構の QKV 生成行列や、その後のフィードフォワード NN 層の重み行列に対して、行列の次元を圧縮することでパラメータ数を削減する手法を考案した。行列のランクを落とす手法のため、LoRA（Low-Rank Adaptation）と呼ばれる。その際、事前学習されたモデルのパラメータは固定し、それと並列する形で次元圧縮されたモデルを並列に追加することにより、事前学習で習得した学習内容を劣化させることなく、新規情報を学習させている（図表 35）。

図表35　LoRAの概念図：左がオリジナルモデル、右が並列設置した次元圧縮モデル



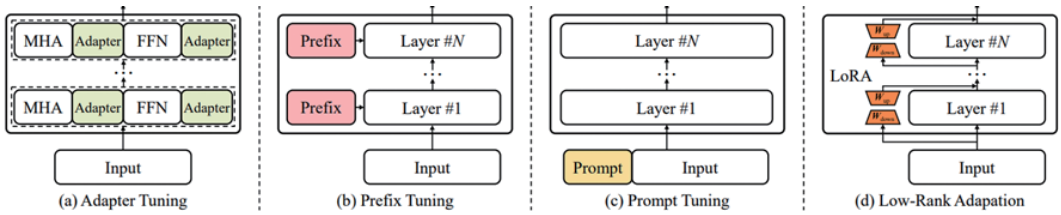
出所）Hu et al.（2021）

Li and Liang（2021）は、Prefix-Tuning と呼ばれる手法を提案した。図表 36（b）がその概念図である。事前学習済みモデルに対する入力値に特定のプレフィックス（接頭語 / 接頭情報）を追加し、そのプレフィックスに対するパラメータ部分のみを学習する（他の入力データにかかるパラメータは不変）。このほか、Selective と呼ばれる手法があり、Transformer の各層にあ

るバイアス（定数項）のみを再学習する手法である。

なお、PEFT といえども、GPT-3 以降のようにモデルが巨大化すると新規に学習するパラメータが増加する。その次元を絞り込み過ぎるとパラメータが過少なことによる性能低下が生じかねない（図表 37 の左下がり部分）。そこで、図表 36（c）にあるような Prompt Tuning も考えられた。プロンプトエンジニアリングと似ているが、タスクに適したプロンプトを自動的に生成・最適化する点が異なる。前者は人間が試行錯誤しながらプロンプトを工夫するアプローチであり、使い方のチューニングに相当する。

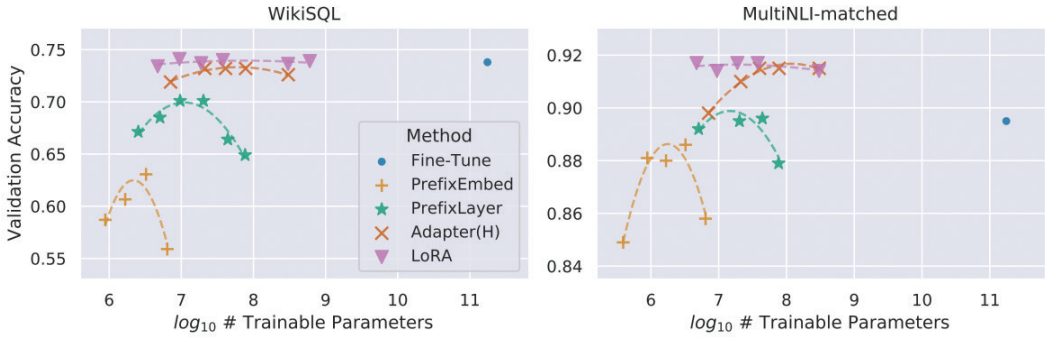
図表36　ファインチューニングのアーキテクチャ



出所）Zhao et al.（2023）

図表 37 は、LoRA を提唱した研究論文が示した PEFT 各手法の比較である。オリジナルモデルをフルにファインチューニングした結果が右側の丸印で（パラメータ数が多い）、各手法についてパラメータ数を変更しながら性能の変化を見たものが各手法別に点線で示されている。LoRA の学習性能の高さと、パラメータを絞り込んでも劣化しにくい性質が観察されている。

図表37　PEFT各手法の性能比較



出所）Hu et al.（2021）

5.4 RAGとPEFTの選択

図表 38 に RAG と PEFT の特徴を対比した。適切な応用分野としては、定型的にはチャットボットや FAQ、ニュース作成、専門分野の診断（医療 / 法 / IT）、内部ナレッジ集積・活用、専門分野翻訳・要約など共通するところが多いが、定性的にみるとサービスの内容に応じて両者を使い分けた方がよいことがわかる。

図表38　RAGとPEFTの比較

	RAG	PEFT
長所	外部知識の活用により広範な情報資源にアクセス可能、情報の更新頻度も操作可能、モデルサイズが小さくとも性能が期待できる	事前学習済みのモデルを効果的に活用、フルパラメータの再学習に比べ破滅的忘却のリスクが低い
短所	知識データベースのメンテナンスが必要、パフォーマンスが検索精度に依存、検索した情報を LLM が上手くまとめきれないことや一般情報のノイズが入る場合もある	適応できるタスクに制限がある場合がある、モデルの対応領域に限界がある
実装	外部知識データベースの構築や、データパイプラインの実装知識が必要	各種 PEFT 固有のモデル追加（アダプター層の追加等）が必要
拡張性	知識ベースの拡張で柔軟に可能	新しいタスク（ドメイン）ごとにファインチューニングが必要
適切な応用分野	広範な知識を必要とし、知識の更新頻度も比較的高い分野	ファインチューニング・追加学習が行いやすい分野、知識領域の限定と低頻度更改

出所）筆者作成

5.5 AIエージェントの活用、LangChainフレームワーク

RAG、PEFT に続く 3 番目の対応は、AI エージェントの活用である。今日の東京の天気をリアルタイムで知りたい場合、天気予報サイトに API 接続して情報を収集し、これをもとに LLM で文章化するほうが確実である。ここでも LLM は自然言語インターフェイスとして機能しており、知識や情報収集は API 先を訪問して情報を収集してくる AI エージェントが担うことになる。

こうしたシステムの構築においては、RAG と同様、LLM は構成要素の一部に過ぎず、AI エージェントを機能させるセンターや、収集情報の受け渡しを担うデータパイプラインの設計が必要となる。以上のような IT システム構築のコストを削減してくれる生成 AI フレームワーク兼ライブラリとして LangChain が 2022 年 10 月にリリースされ、2023 年を通じて急速に普及した。LangChain は、LLM を活用したアプリケーションの開発を支援するためのフレームワークであり、これを活用することで、データ取得、前処理、モデルの呼び出し、出力解析など、複数のステップを統合して一連のワークフローを構築できる。ライブラリとして様々なサービスを備えており、ベクトル化データベースの検索や、エージェントの起動など、種々のタスクに対応するためのモジュールを提供し、開発者がチェーン（処理の流れ）を短時間のうちに構築することを可能にする。直感的な API を提供しており、特定の機能やタスクを呼び出すためのインターフェイスが利用できる。6 節では Python から LangChain や LlamaIndex を使って RAG や AI エージェントを構築した簡単な事例を紹介する。

6．アプリケーション実装の進化と学び方

企業が生成 AI を活用した自社サービスを構築する際には、適切な LLM の選択以外にも種々の課題が伴う。どのような IT 環境で、どのような（LLM 以外の）サービスと連携させて開発するのか、社内情報など非公開情報を利用する場合には安全な環境をどう構築するのか、非公開情報の活用は RAG と PEFT のいずれを使うのか、利用者のアクセスをどう管理するのか（誰が何を知っていいのか / 情報を使っていいのか）、品質管理にかかる問題への対応（例えば情報の更新やサービスシステムのメンテナンス）、サービスの内容や技術の陳腐化に伴う寿命予測（サービス構築の投資判断）、生成 AI サービスの公正性や公平性、倫理の確保、データ保護規制などの法令順守、開発スタイル（内製化か外注か、人材の確保は）といった種々の課題が伴う。更に高い次元の課題としては、企業全体のデジタル経営戦略に生成 AI をどう位置付けて、資金や人材といった経営資源を確保・投入するのか、組織体制をどうデザインするのかといった経営上の論点もある。

これらの課題のうち、本節ではアプリケーション実装にかかる IT インフラの選択や、実装を進めるに際しての学びのプロセスの参考になるような情報を提供する。これらは、IT 戦略すなわち経営戦略と直結する論点となるため、技術論に止まる話ではない。

6.1 段階的学びのステップを通じて知るインフラ技術

学びという視点から、生成 AI アプリケーションサービスの使い方と作り方を順に並べると、例えば以下のような 6 ステップが考えられる。段階的学びの過程において生成 AI アプリケーションサービス構築に必要な IT インフラ知識も高まっていく。

Step1　生成AIアプリケーションサービスの体験

まず、商用（有償 / 無償）の公開サービスを利用して、生成 AI サービスにどのような種類のものがあり、どのような UI で提供され、どのような UX（ユーザー体験）をもたらしているのかを「身をもって知る」ことが第一歩目である。例えば、ChatGPT や Gemini（のチャットサービス[34]）、1 節で紹介した notebookLM や GPTs の各種サービスを利用し、研究者であれば Elicit という論文検索・内容分析ツールや perplexity（出典に基づいて回答を作成する対話型検索エンジン）を使ってみることがあげられる。現在の ChatGPT はエージェント機能がバックエンドで動いており、専門的な質問をするとインターネットで検索した結果を情報源として GPT-4 の言語生成能力と汎用知識をあわせて回答文を作成している。また、notebookLM は RAG の実装事例であり、本稿を読んだ後に利用すると UI の背後で動いている仕組みがイメージできる。notebookLM の裏で動いている LLM は Google の Gemini 1.5 Pro である。これと同様に、各種の生成 AI アプリケーションサービスでは、LLM 自体は前面に出てこず、その機能を使って何をユーザが行いたいか（サービスプロバイダが何を届けたいか）に焦点をあてた作り込み方になっている。

34：GPT シリーズはチャット web アプリケーションサービスを ChatGPT と別名シリーズで提供しているが、Gemini シリーズでは、API サービスを提供する LLM ライン（Gemini 1.5 Pro/Flash など）とチャット web アプリケーションサービスを同名としている。

以上のような生成 AI 入門学習体験だけでも、ここ 1 ～ 2 年の LLM 関連技術の大幅な進歩が体感できる。ChatGPT が初登場してきたころとはサービスの内容や質が劇的に変化している点に未だに気が付いていないビジネスパースンや研究者は少なくない。最新のサービスを使ってみるという最初の一歩が一番重要である。

Step２　Python入門

生成 AI アプリケーションサービスの開発に必要なコンピュータ言語として Python を学ぶ。JavaScript/TypeScript を Node.js 実行環境で動かすアプリケーションを作成する方法に対応している LLM や関連サービスもあるが、Python が最も汎用性が高く、かつ AI / 機械学習全般に活用されているため応用範囲も広い。また、学びにかかる初期コストも比較的小さい。Python を用いると LLM ベンダーなどから提供されている API 接続サービスが簡単に実現できる。curl コマンドを用いてシェルから投げる方法もあるが、学びには不便である。Python の利用環境はインタラクティブなノートブック形式でコードを実行できる Jupyter Notebook や同様なスタイルでクラウドの高性能計算環境を間借りできる Google Colaboratory（Google Colab）がある。このほか、VSCode（Visual Studio Code）のような汎用コードエディタに Python 拡張機能をインストールして使用する方法もある。最初からクラウド上でのアプリケーション構築に進むなら、Azure Notebooks（Jupyter Notebook）環境や AWS の SageMaker 環境を用いる方法もある。

Step3　APIを使ったLLM使用

Python の基本的な使い方が理解できたら、LLM の公式資料や、ネット上にある無料学習教材、各種の体験記事（Qiita や note、Zenn に大量にある）などを参考にしながら、API を経由して商用 LLM サービスを使ってみる。商用 LLM 開発企業は API を通じて LLM サービスを提供している。Step1 で示した生成 AI アプリケーションサービスは UI を各社で作り込んではいるが、その入力情報は高性能の商用 LLM（サービス内容によっては閉鎖環境下のクラウド LLM）に対して API で引き渡され、処理されたものを受け取ってブラウザで提示するといったサービス構成を基本的に採用している。顧客向けサービスを提供する各社が LLM を各々の自社サーバで運用しているわけではない（そうしたケースも存在しうるが、何らかの経済合理性などがある場合に限られるであろう）。

このため、生成 AI アプリケーションサービスでは API の利用が必須となる。その利用法は使いやすく整備されており、アカウントの取得や API-Key の管理などに気を付ければ API 接続は容易に行うことができる。Python などを使って自ら作成した生成 AI アプリケーションサービスから API を通じて外部 LLM を呼び出すことが可能となる。

Step4　LangChain/LlamaIndex/Difyを使う

汎用 LLM を単体で使うだけでは、企業内情報や顧客情報を活用した

サービスは提供できない。外部データを活用するRAGをLangChainやLlamaIndexを用いて作成してみる。これらはオープンソースとして提供されているため、誰でも利用することができる。LangChainやLlamaIndexの公式ドキュメント、ネット上の解説マテリアル、書籍（布留川 2023・2024、田村 2023）などを参照すると簡単に学ぶことができる。LLMの使い方は学べても、アプリケーションサービスに仕立てるには多くの周辺技術の学びと実装が必要となるが、LangChain/LlamaIndexはそのコストを大きく引き下げた。これらは生成AIアプリケーションサービス開発の民主化の象徴的存在となっている。これらを用いるとAIエージェント系のサービスも容易に実装できる。その際、ネット検索に利用するサービスでAPI登録、API-Key発行が必要となるが、簡単に取得できる。

また、ノーコードで生成AIアプリケーションサービスを構築する技術も進歩している。Difyはオープンソースの開発プラットホームで、2024年に入って急速に注目を集めるようになった。クリックだけでRAGシステムほか多様な生成AIアプリケーションサービスを構築することができる。LangChain/LlamaIndexは、便利とはいえコードを書ける技術を必要とした。Difyを始めとして、今後も増えるであろうノーコード開発ツール群は、生成AIの民主化を更に推し進めることになろう。

Step5　クラウドサービスの基本を学ぶ

Step4だけでは、非公開情報を閉鎖環境で運用し、利用権のあるユーザだけに対してサービスを提供することができない。これを実現するITインフラとして、自社サーバにオープンソースのLLMを立てるという選択肢もありえるが、クラウドサービスを使ったセキュアな閉鎖環境やユーザの登録・認証システムの構築が、開発速度や運用の柔軟性、コストの面で現実的である。クラウドベンダー各社は、LLMだけでなくアプリケーションサービスを構築する際に必要となる様々なマネッジドサービスを提供している。これらを用いることで、情報データベースや検索サービス、データパイプライン、UIとなるWebアプリケーションサービスなどを一体活用できる利点がある（図表33参照）。Amazon Bedrockのように様々な最新のLLMを選択可能なかたちで提供するクラウドサービスもある。もちろん、LangChainのようなオープンソースサービスをクラウド上で利用することも考えられるが、本番環境に展開していった際の可用性や統合運用性を考えると、クラウドのマネッジドサービスの利点は大きく、構築したいシステム内容やコスト面も合わせての判断となろう。

Step6　Azure OpenAIやAmazon Bedrock、Google Cloud Vertex AIを使う

パブリッククラウドサービス活用の基本がマスターできると、Azure OpenAIやAmazon Bedrock、Google Cloud Vertex AIなどの統合サービスを用いた生成AIアプリケーションサービスに到達する。ここに至って、課題であった①企業内部情報や顧客情報をどう活用して生成AIアプリケー

ションサービスを構築するか、②その際、情報管理の安全性をどのようにして確保するかが解決可能となる。クラウドベンダーのサイトには、システム構築に必要な情報や学習マテリアルが多く掲載されている。永田他（2024）や御田・熊田・森田（2024）、坂本（2022）ほかクラウドサービス上での生成 AI アプリケーションサービス構築に関する書籍も出版されている。

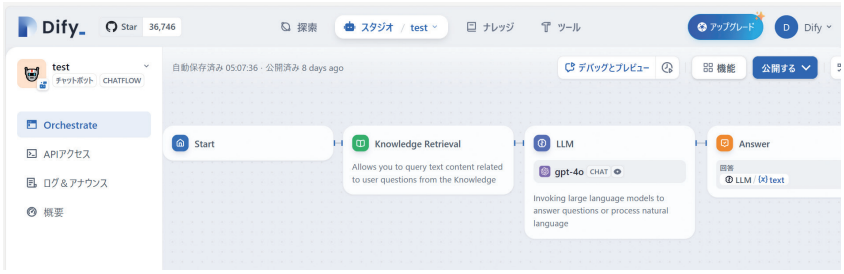
6 つのステップを順に踏む必要はないが、一足飛びにクラウドサービスを契約して開発を進めようとすると、技術的スタックを内部で積み上げていないため、外注に頼りがちとなる。また、PoC での手触り感もないまま進めることになる。生成 AI に限らず、技術進歩が非常に早く、サービスの陳腐化リスクが高い分野では、外注依存のサービス開発体制はマイナス面が大きい。ここに示したのは学び方のステップの一例に過ぎないが、サービスを創造・運営していくビジネスの現場が何らかの形で内製化に関与していく経営に切り替えていく必要がある。これは生成 AI に限らず、IT サービス産業化するデジタル社会でのビジネス全般に共通するテーマである。

6.2 RAG実装の３事例

1 節で利用例として示した notebookLM は RAG を一般向け公開 Web アプリケーションサービスとして実装したものであった。図表 3 のアプリケーション画面左側は、読み込ませたレポート 24 本がリストされており、これらの情報はベクトルデータベース化されている。下の入力欄からの問い合わせに対応して、データベース情報が検索され、適切なものが抜き出され、プロンプトとあわせて Gemini 1.5 Pro に API 経由で送信され、回答を受け取ってブラウザに表示されるという情報処理の流れとなる。

これと同じものをノーコード開発の Dify で簡単に作成することができる。図表 39 に示した Dify のパネルには、問い合わせを受ける「スタート」から右側に向かって「知識データベース」、「LLM（GPT-4o、筆者所有アカウントの API-Key を利用）」、「回答」という構成が取られていることが示されている。図表 40 は情報データベースの中身である。URL 指定で登録されたレポートが並んでいる（図表 3 と同じ）。図表 41 はその中身の一つを見たもので、レポートが細かい部分に分割されて（チャンクされて）格納されていることがわかる。文章はチャンク単位でベクトル化されて保管され、利用時に検索対象となり、対象となった場合はプロンプトと合わせて LLM に送られる。関連する部分を過不足なく選択するため、文章は細かくチャンクされる。図中に示されている情報は、元は html ファイルであるが、パーサ（parser）機能によって不要なタグ情報を削除して格納されている。Dify は Firecrawl という高機能の Web クローラー兼 Markdown 形式変換 API サービスを今年 6 月に組み込んだため、図表 41 のように綺麗な整形物を簡単に得ることができる。URL の下位ディレクトリ指定や排除もメニューから簡便に行える。元情報が PDF であった場合、段落・見出し・図表などを識別し、適切にパーサする能力が重要となる。こうしたパーサの性能も RAG を支える重要な構成要素となっている。

図表39　Difyのメインパネル



出所）筆者作成

図表40　Difyのナレッジパネル

出所）筆者作成

図表41　htmlファイルからの格納状態:チャンク分割

出所）筆者作成

最後に、Python を用いて LangChain を使った RAG システム構築の事例を紹介する。図表 42 はその一部を示しており、まず Tavily という AI エージェント専用に構築された検索エンジンを使って、ブロックチェーンを使った資産のトークン化に関する質問に関連した情報をインターネット上から収集してくるよう AI エージェントに命令している。具体的には、RWA（Real World Asset）のトークナイゼーションに際して日本のどのような法律が関

係してくるかを問うている。図中には収集した結果を示しており、ネット上にあった異なる pdf ファイルから、関連度が高そうな箇所をパーサして整えたうえで、収集していることがわかる。

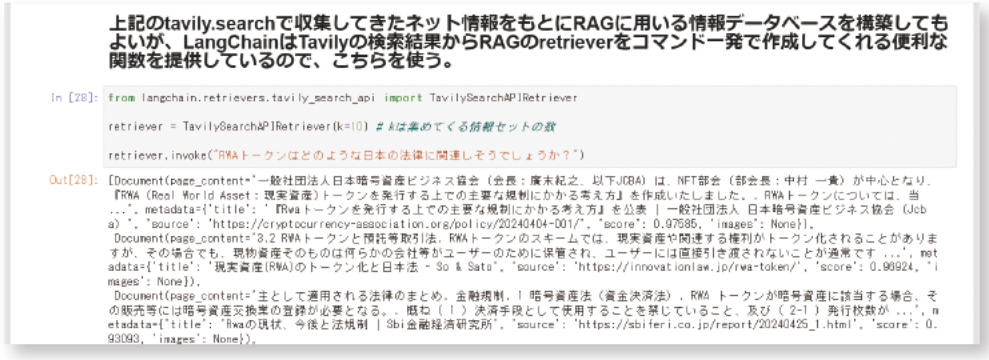
図表42　Webサーチ用エージェント

出所）筆者作成

図表 43 では、収集した情報から情報データベースを作って検索装置を繋ぐという標準的な手続きは行わず、一気に RAG を構築する手法を採用している。TavilySearchAPIRetriever で RAG オブジェクトを作成し、そこに web サーチの結果をストアするコードを示している。これがライブラリの読み出しと生成オブジェクトの利用を含めて僅か 3 行で完結する。3 行目では、同オブジェクトの呼び出し（invoke）機能から質問内容を読み込んで、回答を生成している。図表 44 では、この情報データベースからの検索結果を LLM（ここでは GPT-4o のチャットシステム）に与えて、文章を生成させたものを表示している。プロンプトでは、提供された情報のみに基づいて回答するよう指示を与えている。これも 10 行程度のコードで実現できている。

このように LangChain を利用すると、簡単に情報データ指定型の RAG や web サーチ・エージェント型の RAG を構築することができる。また、PEFT で LoRA を試してみたい場合は、HuggingFace という AI 技術のライブラリから Python に読み込むことができる。この場合は計算資源を必要とするので、ローカル PC の Python 環境ではなく Google Colab を利用した方がよい。

図表43　サーチ結果からのRAGオブジェクト生成とその使用



出所）筆者作成

図表44　RAGを使ったチャットシステム

出所）筆者作成

7．おわりに

本稿では、生成 AI ウォークスルーと題して、ニューラル言語モデルの基礎、現在の LLM に繋がるモデル発展、様々な LLM の群雄割拠、発展の過程で発見されてきた多様な転移学習の新形態、それらがもたらした利用法の拡大、RAG や PEFT の技術、実装に必要な技術群とその学び方、実際の RAG 構築事例を紹介した。こうした情報は、生成 AI 技術を金融実務に活用していく際の学びや実践の水先案内（Pilot）となろう。6 節で紹介した Step1 からの実践を通じて、生成 AI のビジネス活用が決して難しいものではないことが体感できると思う。また、筆者が参加した証券アナリスト協会主催の座談会（和泉他 2024）では、金融機関が生成 AI 活用を推進する際の様々な課題が議論されており、こちらも参考となろう。

この分野は、過去数年のうちに急速に発展したため、ほぼ全員が新規参入者として同じスタートラインにいる。もちろん過去の技術スタックの有無はロ

ケットスタートの実現に影響するが、技術の民主化によりハードルは大きく引き下げられた。6 節で紹介したノーコード開発が典型例である。全力で走り出したものから市場や付加価値（利益）の創造に近づいていけることは間違いない。本稿がその Pilot となれば幸甚である。

最後に、自然言語処理技術の可能性について 1 点触れたい。SBI 金融経済研究所の所報 5 号で水田（2024）は、人工的な金融市場を用いたシミュレーションの可能性について様々な研究を紹介している。人工市場の研究においては、エージェントの振る舞いをどう現実的に設定するか、どこまで複雑な振る舞いをモデル化できるかが、結果の妥当性やリアリティを担保する鍵となる。生成 AI は膨大なコーパスとそこに含意された人間の思考パターンを巨大なモデルとパラメータ群で学習している。これを人工市場のエージェントとして用いることには大きな可能性があると考えられる。

類似の発想が Ha and Schmidhuber（2018）によって World Models として提唱されており、強化学習やロボティクス、AI シミュレーションの分野で大きな影響を与えている[35]。World Models では、エージェントが環境の内部モデル（ワールドモデル）を構築し、観察からの学習によって「ワールド」の状態や状態の推移則を認知する。そのうえで、潜在的な報酬を最大化するように自らの行動を決定する。最適化のスコープのバリエーションは色々ありえて、例えば、エージェントが行動を履行することがワールドへ何がしかの影響を及ぼし、これが自己へフィードバックしてくる点も考慮したうえで行動選択を決定するという設定も考えられる。

同論文では、仮想空間上での車の自動運転学習が取り上げられているが、モダンマクロ経済学の DSGE モデルを学んだものは、発想の親和性の高さと DSGE の限界を意識するであろう。すなわち、エージェントである代表的個人は完全に動学最適化を解いたうえで行動を選択する（未来にわたる動学的パスを含めて決定し、毎期生じる外生ショックに対して最適化計算をやり直す）という合理的個人の仮定とその限界である。この仮定を緩めるとアドホックなモデルがいくらでも構築でき、観測された現実が仮定によって自在に説明できてしまう。上述した人工市場の研究におけるエージェントの振る舞い設定の適切さに関わる難しさと同根の問題である。

しかし、人間の認知や意思決定の複雑さにかかる部分が重要であることは、近年の非伝統的金融政策の振り返りで話題に取り上げられている「デフレのノルム」の議論においても確認される。財サービス価格を引き上げると市場シェアを失うため、生産コストの抑制、とりわけ賃金の抑制で対応しよう（幸いデフレであり雇用者との賃金交渉もそれを可能にしているほか、労働者にとってもデフレならば賃金据え置きは容認可能）といった「ワールド（世界の在りよう）」の認知が企業側にも労働者側にも成立し、すなわちデフレのノルム（社会通念）が定着し、自己実現的な均衡に陥ったというものである。こうした均衡が成立し、長期間継続してしまった理由を検討し、どのような外生ショックやメカニズムがデフレ均衡からの離脱をもたらすのかという検証を行う際に、ピュアな DSGE モデルの適用には限界がある。

経済環境や金融市場というワールドを人間がどう認知し、それを行動に反

35：David Ha は、元 Google Brain で現在は Sakana AI の CEO である。Jürgen Schmidhuber は、本稿で紹介した LSTM の考案者の一人である。ちなみに、Sakana AI の CTO である Llion Jones は、Transformer モデルの考案チームのメンバーである。

映させるかを考える際に、World Models の考え方は一つのヒントになろう。金田・坂地（2023）は、気候変動の原因と結果に関するナラティブがどのように成立したか、株式市場がこうしたナラティブにどう反応したかを BERT やテキストからの因果抽出技法によって検証している。認知科学と自然言語処理と金融経済学の接点であろう。こうしたワールドモデルが繋ぐ学際的アプローチは、Web3 がイメージする分散分権型の仮想社会で、エージェントがどう振る舞い、その集合体としてのシステムがどう振る舞うかを検証するうえでも有益なツールとなりそうである[36]。LLM をはじめとする AI 技術の応用範囲は広い。

36：OpenAI が 2024 年 2 月に公表した Sora は、テキストプロンプトから動画を自動生成する生成 AI モデルとして注目を集めた。黒いレザージャケットとサングラスを身につけた女性がアジア風の夜の繁華街の濡れた道路を歩く有名な動画である。メディアでは動画生成 AI として取り上げられたが、OpenAI の website は、"Video generation models as world simulators" と紹介しており、サイトの URL もそうネーミングされている。Sora は世界モデルの構築技術の一つであり、OpenAI は「Sora は AGI を達成するための重要なマイルストーン」と述べている。

参考文献

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014), "Neural machine translation by jointly learning to align and translate," arXiv:1409.0473, 2014.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent and Christian Jauvin (2003), "A neural probabilistic language model," *Journal of Machine Learning Research*, Vol.3, pp.1137-55, February 2003.

Bengio, Yoshua, Réjean Ducharme and Pascal Vincent (2000), "A neural probabilistic language model," *Advances in neural information processing systems*, vol.13, 2000.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword Information," arXiv:1607.04606, 15 July 2016.

Brown, Tom, et al. (2020), "Language models are few-shot learners," Advances in neural information processing systems 33 (NeurIPS 2022), pp.1877-1901, 2020.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio (2014a), "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," arXiv:1406.1078, 2014.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau and Yoshua Bengio (2014b), "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," arXiv:1409.1259, 2014.

Clark, Kevin, Urvashi Khandelwal, Omer Levy and Christopher D. Manninget (2019), "What does BERT look at? An analysis of BERT' s attention," arXiv:1906.04341, 2019.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805(v1), 11 October 2018.

Elman, L. Jeffrey, (1990), "Finding structure in time," *Cognitive Science* vol.14, 2, pp.179-211. March 1990.

Feng, Junxi, et al. (2019), "Reconstruction of porous media from extremely limited information using conditional generative adversarial networks," *PHYSICAL REVIEW E* 100, 2019

Gunasekar, Suriya et al. (2023), "Textbooks Are All You Need," arXiv: 2306.11644(v2), 2 October 2023.

Ha, David and Jürgen Schmidhuber (2018), "World Models," arXiv: 1803.10122(v4), 9 May 2018.

Harsoor, Sharan (2023), "Transformer model and variants of transformer (ChatGPT)," Medium, 7 June 2023. https://pub.aimind.so/transformer-model-and-variants-of-

transformer-chatgpt-3d423676e29c
Hendrycks, Dan, et al. (2021), “Measuring massive multitask language understanding.” arXiv:2009.03300(v3), 12 January 2021. (Note this is not the version1 in 2020.)
Hochreiter, Sepp, and Jürgen Schmidhuber (1997), “Long short-term memory,” *Neural computation 9.8*, pp.1735-80, 1997.
Hoffman, Jordan, et al. (2022), “Training Compute-Optimal Large Language Models,” arXiv: 2203.15556(v1), 29 March 2022.
Hopfield, John (1982), “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the national academy of sciences 79.8* pp.2554-58, 1982.
Houlsby, Neil, et al. (2019), “Parameter-efficient transfer learning for NLP.” International conference on machine learning, PMLR, arXiv:1902.00751(v2), 13 Jun 2019.
Hu, Edward J., et al. (2021), “Lora: Low-rank adaptation of large language models.” arXiv:2106.09685(v1), 2021.
Jordan, I. Michael, “Serial order: A parallel distributed processing approach,” Institute for Cognitive Science, University of California, San Diego, ICS Report 1986, May 1986.
Kaplan, Jared, et al. (2020), “Scaling laws for neural language models.” arXiv:2001.08361, 23 January 2020.
Karpathy, Andrej (2015), "The unreasonable effectiveness of recurrent neural networks," Andrej Karpathy Blog, posted on May 21, 2015.
Le, V. Quoc and Tomas Mikolov, “Distributed representations of sentences and documents,” arXiv:1405.4053, 16 May 2014.
Li, Xiang Lisa and Percy Liang (2021), “Prefix-Tuning: Optimizing Continuous Prompts for Generation,” arXiv:2101.00190(v1), 1 Jan 2021.
Patrick Lewis et al. (2020), “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” arXiv: 2005.11401(v1), 22 May 2020.
Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain and Jianfeng Gao (2024), “Large language models: A survey.” arXiv:2402.06196, 2024.
Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur, “Recurrent neural network based language model,” *Interspeech* 2010, pp.1045-48, 26 September 2010.
Open AI et al. (2023), “GPT-4 Technical Report” , arXiv:2303.08774(v1), 15 March 2023. (Latest version is 6th.)
Penedo, Guilherme, et al. (2023), “The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only,” *Advances in Neural Information Processing Systems* 36, 2023.
Peters, E. Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer (2018), “Deep contextualized word representations,” arXiv:1802.05365, 15 February 2018.
Radford, Alec, Karthik Narasimhan, Tim Salimans and Ilya Sutskever (2018), “Improving language understanding by generative pre-training,” 2018.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever (2019), “Language models are unsupervised multitask learners.” OpenAI blog 1(8),9, 2019.
Rothman, Denis (2021), Transformers for Natural Language Processing, Packt Publishing, 2021.（デニス・ロスマン（2022）、『Transformer による自然言語処理』、黒川利明訳、朝倉書店）

Ruan, Yangjun, Chris J. Maddison and Tatsunori Hashimoto (2024), “Observational scaling laws and the predictability of language model performance.” arXiv:2405.10938(v1), 17 May 2024.
Sardana, Nikhil and Jonathan Frankle (2023), “Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws,” arXiv:2401.00448(v1) 31 Dec 2023.
Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo (2024), “Are emergent abilities of large language models a mirage?,” *Advances in Neural Information Processing Systems* 36, 2024.
Sutskever, Ilya, Oriol Vinyals and Quoc V. Le, “Sequence to Sequence Learning with Neural Networks,” Advances in neural information processing systems, 2014.
Thompson, D. Alan (2024), “Chinchilla data-optimal scaling laws: In plain English,” LifeArchitect.ai, February 2023, updated Jun 2024. https://lifearchitect.ai/chinchilla/
Vasnetsov, Andrey (2024), “BM42: New Baseline for Hybrid Search,” Qdrant web site, 1 July 2024, https://qdrant.tech/articles/bm42/
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, “Attention is All you Need,” *31st Conference on Neural Information Processing Systems*, NIPS 2017.
Wei, Jason et al.(2022), “Emergent abilities of large language models,” arXiv:2206.07682(v2), 26 October 2022.
Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin and Xia Huet (2024), “Harnessing the power of LMs in practice: A survey on ChatGPT and beyond,” *ACM Transactions on Knowledge Discovery from Data* 18.6, pp.1-32, 2024.
Zhao, Wayne Xin, et al. (2023), “A survey of large language models,” arXiv :2303.18223 2023.
Zheng, Huiting, Jiabin Yuan and Long Chen, “Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation,” *Energies*, 10(8), 8 August 2017.
和泉潔他（2024）、「生成 AI が変える金融市場・資産運用業界への影響（座談会）」、証券アナリストジャーナル 62 巻第 2 号、pp.54-78、日本証券アナリスト協会、2024 年 2 月
今泉允聡（2021）、『深層学習の原理に迫る：数学の挑戦』、岩波科学ライブラリー 303、岩波書店
Weights and Biases Japan (2024)、「Nejumi LLM リーダーボード 3 開発の経緯とその評価から見えてきたこと」、note、2024 年 6 月 30 日、https://note.com/wandb_jp/n/nd4e54c2020ce
打田智子・古澤智裕・大谷純・加藤涼・鈴木翔吾・河野普策（2022）、『検索システム：実務者のための開発改善ガイドブック』、ラムダノート
宇根正志、「機械学習による予測・推論の公平性：金融サービスにおいて求められる配慮とは」、金融研究、第 43 巻第 1 号、日本銀行金融研究所、2024 年 1 月
岡崎直観・荒瀬由紀・鈴木潤・鶴岡慶雅・宮尾祐介（2022）、『自然言語処理の基礎』、オーム社
岡崎直観他、「Swallow コーパス：日本語大規模ウェブコーパス」、言語処理学会第 30 回年次大会発表論文集、2024 年 3 月
御田稔・熊田寛・森田和明（2024）、『Amazon Bedrock 生成 AI アプリ開発入門』、SB クリエイティブ、2024 年 6 月
金田規靖・坂地泰紀（2023）、「BERT と因果抽出を用いた気候変動ナラティブの可視化・指数化」、日本銀行金融研究所、ディスカッションペーパーシリーズ 2023-J-7、2023 年 6 月
神嶌敏弘編・人工知能学会監修（2015）、『深層学習』、近代科学社
斎藤康毅（2018）、『ゼロから作る Deep Learning ② 自然言語処理編』、オライリー・ジャパン
坂本俊之（2022）、『Vertex AI で作る AI パイプライン入門』、C & R 研究所、2022 年 11 月
ストックマーク株式会社編（2021）、『BERT による自然言語処理入門：Transformers を使った

実践プログラミング』、オーム社
副島豊（2024）、「金融システムの未来像を探る中央銀行の挑戦」、SBI 金融経済研究所、所報第５号、2024 年３月
――― （1996）、「ニューラルネットワークアプローチによる経済分析：モデルの概要と金融政策への応用例」、金融研究、第 15 巻第３号、日本銀行金融研究所、1996 年８月
田村悠（2023）、『LangChain 完全入門：生成 AI アプリケーション開発がはかどる大規模言語モデルの操り方』、インプレス社
坪井裕太・海野裕也・鈴木潤（2017）、『深層学習による自然言語処理』、機械学習プロフェッショナルシリーズ、講談社
永田祥平他（2024）、『Azure OpenAI ではじめる ChatGPT/LLM システム構築入門』、技術評論社
布留川英一（2023）、『OpenAI GPT-4/ChatGPT/LangChain 人工知能プログラミング実践入門』、ボーンデジタル
布留川英一（2024）、『Google Gemini 1.5 ／ LlamaIndex ／ LangChain 人工知能プログラミング実践入門』、ボーンデジタル
松井孝太・熊谷亘（2024）、『転移学習』、機械学習プロフェッショナルシリーズ、講談社
水田孝信（2024）、「人工市場：金融市場のコンピュータ・シミュレーション」、SBI 金融経済研究所、所報第５号、2024 年３月

次世代金融インフラの構築を考えるに当たっての指針（2024年7月5日公表）

次世代金融インフラの構築を考える研究会

「次世代金融インフラの構築を考えるに当たっての指針」の公表

問題意識

デジタル化社会の進展に伴い、金融サービスの提供主体や提供手段に大きな変化が生じている。金融 API やブロックチェーン技術、ビッグデータの生成と活用に代表されるような情報技術の革新は、新しい決済・送金手段、暗号資産などのデジタル金融資産、分散型金融サービス (DeFi) などの登場をもたらし、金融サービスの内容にも変化が生じつつある。さらには、付加価値（収益）の源泉がデータにシフトし、情報生産機能の高度化が金融機関の経営課題となっている。

こうしたもとで、銀行・証券会社・保険会社などの仲介業者を通じて金融サービスを提供していた現行の枠組みや、中央銀行と民間銀行が提供していた２階層型の決済制度も影響を受けており、セキュリティトークンに係る決済手段やクロスボーダー資金決済の新しいスキームを求める動きの中で、中央銀行デジタル通貨（CBDC）の模索が続いている。

これまで安定運営されてきた従前の金融インフラ（法規制、IT システム、会計ルール、ガバナンス、リスクマネジメント、国際協調など）は、デジタル金融資産の登場によって生じている急激な環境変化に対して十分に適応できているとは言い難く、既存の枠組みの中で対処療法的に変更を重ねることの限界も意識されるようになってきた。例えば、金融仲介業者を通さず転々流通するパーミッションレス型の DeFi に対しても、法定通貨や預金などの伝統的金融資産との仲介機能を果たす金融業者を通じて規制する従前の枠組みを維持していることがその一例である。新しい金融サービスは新しい金融インフラを必要としており、まずはその将来像を描くことが求められている。

設立経緯・研究会メンバー

SBI 金融経済研究所は、デジタル金融の普及によって生じてきている制度的な不適合やデジタル技術のポテンシャルを生かし切れていないために生じてい

る金融サービスの相対的劣化などを明らかにした上で、将来の金融インフラとして望ましい姿を検討し、これを提言としてまとめることによって、社会的な議論を惹起することを目指して、昨年末に「次世代金融インフラの構築を考える研究会」を立ち上げた。本研究会には、決済・金融法制・IT システムの各分野や銀行・証券・暗号資産の業界において活躍されている有識者 6 名をメンバーとして迎え、上述した問題意識のもと議論を重ねてきた。

また、金融庁、日本銀行 (金融研究所) におかれては、オブザーバーとしての参加であり、今回、取りまとめた指針についてはその責を負っていない。

これまでの検討経緯

本研究会では、「暗号資産を初めとするデジタル金融の制度的な枠組みの再構築」をテーマとして取り上げ、金融仲介業者を規制する現行の枠組みを根本的に見直して、主としてデジタル金融資産を対象とした新たな枠組みを模索するべく、2023 年 12 月 25 日の第 1 回会合を皮切りに本年 (2024 年)7 月 2 日の第 7 回会合まで議論を重ねてきた。まずは、国内外の金融サービス利用者、当局を含む金融システムを構成する全ての経済主体に対して次世代の金融インフラの構築を考えるに当たっての指針を示すことが重要であるとの認識のもと、研究会メンバーからのヒアリングや意見交換等を通じて示された視点・留意事項等を別紙の通り「指針」として取りまとめた。

本研究会では、「金融・非金融ビジネスの連携・融合、情報生産機能の高度化、クロスボーダー化に伴う新たな社会的な要請への対応、CBDC を含む新しいマネーシステムの模索といった金融システムの転換期に適応できる次世代金融インフラを構築するため、金融サービスとこれを支える金融インフラ、各々の提供主体など、広範な金融産業構造を含む金融システムを再設計し、これによって国内外の利用者から選ばれる金融システム・金融センターも目指す」という目的設定が重要であるとされた。この目的を達成するためには、従来の経済主体別思考から金融機能別思考に転換するとともに、金融インフラの基盤の上にさまざまな金融サービスが階層（レイヤー）構造として展開され、かつ、それらが有機的に機能するためには階層間の相互依存関係を深く理解することが重要であるとの認識を得た。また、従前の仕組みが内包する課題を解決するためには、これまでとは全く異なる新たな仕組みを導入する必要があり、その際、非効率かもしれないものの、２つの金融インフラを並走させる方策が有効であるとの認識も得た。こうした認識に基づき、「新しい金融インフラの構築を考えるに当たっての指針」を３つに分けて整理するとともに、当面の課題・懸念事項を示した。

(A) 新しい金融インフラの構築に当たって必要となる視点（10 項目）
(B) 新しい金融インフラの構築を考えるに当たっての留意事項（３項目）
(C) 新しい金融インフラを構築する際の進め方（６項目）
(D) 当面の課題・懸念事項（６項目）

また、最後に、課題解決型の思考だけではなく、規範的な思考、すなわち望ましい姿や未来像を想像し、そこに至るプロセスを逆算的に考えるアプローチも取り入れること、変化を続ける環境に対応していくために柔軟性を重視した法規制とすることなどの留意事項を示している。

今後の進め方

本年後半には、これらの指針に照らしつつ、次世代金融インフラに求められる機能等について検討し、2025 年初めにはその大枠について提言をまとめることとしている。

（別紙）

「次世代金融インフラの構築を考えるに当たっての指針」の概要

1．次世代金融インフラを考える目的とその視座

（A）次世代金融インフラを考える目的

（B）次世代金融インフラを考えるに当たっての視座

2．「我が国が目指すべき金融システム」の実現を可能とする金融インフラを考えるに当たっての指針

（A）新しい金融インフラの構築に当たって必要となる視点（10 項目）

① イノベーションの促進・活用
② 市場の公正性 (インテグリティ) の確保
③ セキュリティ（安定性・安全性）の確保
④ 柔軟性のある仕組み
⑤ 適切なガバナンス構造・所有構造の検討
⑥ 業態間・国内金融インフラのボーダーレス化への対応
⑦ 国際競争の激化への対応
⑧ デジタル化に対応した法規制・監督体制のあり方
⑨ 経済安全保障への対応
⑩ DX 推進に対する官民による積極的な取組み：協調と競争

（B）新しい金融インフラの構築を考えるに当たっての留意事項（３項目）

① 金融分野の活動（金流）と非金融分野の活動（商流）の連携・融合が産み出す価値
② 金融サービスと金融インフラの相互依存
③ ブロックチェーン技術・DLT（分散型台帳技術）・トークン化、ビッグデータ・AI などを利用したデジタル金融

（C）新しい金融インフラを構築する際の進め方（６項目）

① 標準化と統合台帳（Unified Ledger / Shared Ledger）の動きへの対応
② 標準化に向けた当局の積極的な関与
③ 情報収集・分析能力の程度による投資家の区分け
④ リテール金融とホールセール金融の区分
⑤ ニーズ志向とシーズ志向
⑥ 規模・属性などに基づく段階的な規制の活用

（D）当面の課題・懸念事項（６項目）

① セイフティネットのあり方
② リスクに応じたコンプライアンス対応
③ 金融機関グループ内における情報共有の阻害要因（ファイアウォール規

制など）への対応
④ 金融機関による DX 人材の育成・確保
⑤ 我が国の法体系（資金決済法・金融商品取引法・銀行法等）の整理・拡充と金融制度のあり方
⑥ 暗号資産などのデジタル金融資産の投資該当判断基準の国際標準化

3．提言に当たって留意すべき事項

① 分かり易く、柔軟性に富んだ提言
② 全く新しい法規制・制度を構築する際の視点
③ 短期的な視点と中長期的な視点

次世代金融インフラの構築を考えるに当たっての指針

1．次世代金融インフラを考える目的とその視座

（A）次世代金融インフラを考える目的

- 経済社会のデジタル化が進展する中、金融・非金融ビジネスの連携・融合、情報生産機能の高度化、クロスボーダー化に伴う新たな社会的要請への対応、CBDC を含む新しいマネーシステムの模索といった金融システムの転換期を迎えており、これらの諸問題への対応は待ったなしの状況にある。
- 本研究会では、金融インフラを「法規制や制度、IT インフラ、金融サービス・インフラの産業構造、金融システムのガバナンスを含む広範な概念」として捉え、経済成長や国民厚生の向上を促す金融サービスの創造や金融産業構造の改善をもたらし、金融イノベーションの推進に資する次世代の金融インフラを構築する際の指針をまとめた。今後、ここで示した指針に基づき、現在直面している諸課題への対応方針や望ましい金融システム像の実現に向けた道筋を俯瞰的な視点で考えつつ、金融システムの再設計を模索し、次世代金融インフラの将来像について提言をまとめることとしている。本指針が、国内外の金融サービス利用者、当局を含む金融システムを構成する全ての経済主体による諸課題の解決に向けた取組みに資することを期待する。
- これにより、アジア圏の金融ハブ機能を担うことも含めて、国内外の利用者から選ばれるグローバルな金融システム・金融センターも目指す。なお、日本の経済成長・国民厚生の向上のための金融システムと、グローバル経済を相手に付加価値・収益を生み出していく金融システムという 2 つには相互補完性が存在している。

（B）次世代金融インフラを考えるに当たっての視座

- 上述した目的を達成するためには、従来の経済主体別思考から金融機能別思考に転換するとともに、第 2 節に指針として示した多様な視点・留意事項を考慮しつつ、デジタル化社会の到来によってポテンシャルが大きく拡大した情報の利活用を通じた金融イノベーションの推進を念頭に置くことが求められる。その際、現代のＩＴシステムの基本的な設計思想である階層構造（機能がレイヤー状に実装されていること）を念頭に、金融インフラと金融サービスの間や、金融インフラ内部に存在する階層構造や相互依存構造を深く理解した上で、高度に機能する金融インフラを構築することが重要である。併せて、新たな金融インフラの構築を機に、デジタル化に対応した金融制度（法規制・監督体制など）のあ

り方を検討する必要がある。

- 従前の仕組みが内包する課題を解決するためには、デジタル金融を活用しつつ、これまでとは全く異なる新たな仕組みを導入する必要がある。その際、従前の仕組みを今の金融システムを担う基盤として維持しつつ、新たな仕組みを模索・創造するためには、２つの金融インフラを並走させる方策が有効である。この場合、現行の仕組みに縛られないためにも、新たな仕組みを導入する際、場合によっては従前の運営主体とは異なる主体とする必要がある。
- ２つの金融インフラの並走は非効率ではあるものの、イノベーションは古いものを凌駕する形で生まれてくるものであり、結局は無駄に終わっても、これを許容しないと新しいものは生まれてこず、社会全体が従前の金融システムに依存し続けることになる。従前の金融インフラの更新版と新機軸なものが併存し、競争し合うことで金融システム全体の改革が成し遂げられると考えるべきである。

  一方で、従前の金融インフラのうち、変えるべきもの、変えることのできないものを見極める必要がある。その際、金融インフラの改善の可能性やその実現の不確実性と整備に係る投資コストを含む多様な視点を考慮する必要がある。

２.「我が国が目指すべき金融システム」の実現を可能とする金融インフラを考えるに当たっての指針

（A）新しい金融インフラの構築に当たって必要となる視点

① イノベーションの促進・活用

② 市場の公正性（インテグリティ）の確保

③ セキュリティ（安定性・安全性）の確保

→ 市場の公正性の確保やセキュリティ（安定性・安全性）の確保など、多角的な視点を考慮しつつ、金融サービスのイノベーション（新たな金融サービスの創造）を最大限引き出す金融インフラを構築する必要。なお、市場の公正性としては、例えば、利用者保護・AML／KYC／CFT の確保、利用者情報の適切な取扱い（プライバシー保護）、知的財産権の保護などが挙げられる。

→ ①のイノベーションを活用して、②③に示す諸点（利用者保護等）の高度化を図る視点が必要。①と②③は、トレードオフの関係にあるものではなく、①は②③の解決手段にもなっている点を念頭に置くべき。

（例）ブロックチェーン技術やプライバシー保護技術の活用も一例。トレーサビリティが確保されているトークナイゼーション、取引台帳やアカウント台帳からの所有者情報の分離と保護など。

→ イノベーションの促進により、資金調達・運用両面において新たな金融サービスを提供できるようにするとともに、金融市場の利用者の意識

も変化し、使い勝手のよい金融市場となることを期待する。
（例）スタートアップ企業やミドルリスク企業への十分な資金供給など。

④ **柔軟性のある仕組み**

→ 将来の技術進歩や全てのリスクを予見することは難しいことから、事前に全ての問題発生を抑制するという考え方から脱却し、柔軟性があり、トラブルに即応できる仕組みとすることが重要。

⑤ **適切なガバナンス構造・所有構造の検討**

→ 新しい金融インフラの構築に当たっては、金融システム全体をどうガバナンスしていくかという視点が重要。これには、金融システムを構成する経済主体、すなわち金融機関や顧客、金融インフラ提供主体、規制当局などの適切なガバナンス構造や所有構造を検討し、これらの間の相互関係、特にインセンティブ構造を理解する必要がある。

→ 金融インフラ運営主体のガバナンス構造の検討に当たっては、IOSCO の示す報告・勧告も参照しつつ、市場の公正性や投資家保護の達成を目指して、リスク評価、包括的な開示、情報共有などを図る必要。

⑥ **業態間・国内金融インフラのボーダーレス化への対応**

→ デジタル化の進展に伴い、国内金融市場では業態を超えて競争が激化する方向にあり、一層の効率化が求められている。

→ 従前の銀証保という業態や金融機関・非金融機関という概念が意味をなくす分野が拡大していく可能性が高く、この点を考慮して金融システム・金融インフラを構築する必要。経済主体別の思考から、金融機能別の思考への転換を可能にする金融システムのレイヤー化を図る必要（機能コンポーネントをレイヤー状に自在に組み合わせているＩＴシステムの設計思想への接近）。

→ 従来の金融インフラは、例えば取引所・清算機関・保管振替機構が国内で垂直にサイロ化する制度設計であった。しかし、取引所の合従連衡からボーダーレス化が始まり、新しい台帳技術の登場などにより金融インフラの構造も変化しつつある。このような金融インフラ産業構造の組換えという視点からセキュリティトークン、電子マネー、CBDC、暗号資産、ステーブルコイン、NFT、DeFi、Unified Ledger（統合台帳）、クロスボーダー決済インフラ等を巡る動向を注視することも必要。

⑦ **国際競争の激化への対応**

→ 国際取引のボーダーレス化が進む中、各国の金融市場が国際的に競争し合う方向に進んでいる。誰が金融市場等の運営者になるか、世界の金融機関を引き付けるかという国際競争には各国金融市場の生き残りという側面と相互活用の側面がある。特に、後者の場合は他国の金融市場へのアクセスの可否が重要。

同時に、デジタル化の進展に伴い、各国金融市場間の壁はさらに低くなり、国別（市場間）競争という概念自体が意味をなくす可能性がある。現に、ユーロ建てデリバティブの清算機関がロンドンに存在することに対するユーロ圏の反発はグローバル金融危機のころから生じており、その後のブリグジットが金融センター覇権争いを加速させていることを念頭に、我が国の対応方針を検討しておく必要。

→　対内投資の呼込みや（潤沢な国内貯蓄を活用した）対外投資の活性化に加えて、デジタル化によるボーダーレス化を見据えて、規模と範囲の経済を働かせるため、成長の著しいアジア経済を呼び込むなど、国際的に金融センターとしての地位を確固たるものとするとともに、とりわけアジア圏の金融ハブ機能を果たせることが重要。

→　一方で、内需が大きい日本は、香港やシンガポールなどの戦略とは異なり、日本の実体経済を活性化することが金融市場の発展にも繋がるという点も考慮する必要。

⑧　デジタル化に対応した法規制・監督体制のあり方

→　金融サービスのデジタル化の進展に伴い、法律行為のデジタル化を図る方向で法規制・監督体制のあり方を見直す必要。

（例）監督する側・される側の両方ともにスプテック（Supervisory +Technology）・レグテック（Regulation + Technology）の推進などが求められている。

⑨　経済安全保障への対応

→　金融インフラの構築に当たっては、ボーダーレス化の進展を踏まえ、経済安全保障上の問題が生じないよう対応する必要。市場や台帳を支えるインフラが物理的に存在する拠点や通信エネルギー源、監督権限は、国際政治上の重大な脅威となりうる。一方で、市場重視の自由経済原則については堅持する必要があり、これを脅かす動きにも配慮する必要。

⑩　DX 推進に対する官民による積極的な取組み：協調と競争

→　企業や金融機関、行政の DX 推進は、金融サービスの価値向上や金融・非金融サービスの融合による高付加価値化・高収益化にとって必要条件。

→　次世代金融インフラの構築のためには官民が DX に積極的に取り組む必要。人材育成や　　R & D 投資、高等教育といった社会資源の再配分のみならず、イノベーション促進のため新たな企業文化の創造が必要。官主導で協調する分野と民間による競争する分野を選別の上、多層構造での取組みを行うことも一案。

（B）新しい金融インフラの構築を考えるに当たっての留意事項

① 金融分野の活動（金流）と非金融分野の活動（商流）の連携・融合が産み出す価値

→ 実物経済の課題を解決する手段となり得る金融サービスを目指す必要。そのためにも、金融分野の活動や金融インフラだけを念頭に置いて検討しては不十分。

→ Web3.0 に代表される通り、インターネットの進化により経済活動のデジタル空間・サイバー空間への移行はさらに進展すると考えられる。サイバー空間を場として両分野の連携・融合を推進するため、詳細かつリアルタイムでの両分野間での情報を共有するためのデータの標準化やデータ連携基盤の整備が必要。

→ 特に、金融機関においては、情報生産機能を十分果たして、率先して両分野の連携・融合を図る必要。その場合、必要に応じて銀行による商業への参入が制限されている業務範囲規制などの見直しも検討。

→ 今後は、両分野の連携・融合にこそ付加価値（収益、成長）の源泉があるとの認識のもと、金融を含めたあらゆる産業が金融産業化、IT 産業化していく必要があることに留意するとともに、速やかに実践に移す必要。

→ 必要に応じてデータの目的外利用に係る許諾ルールについても、情報共有を推進する観点から見直すことも重要。

② 金融サービスと金融インフラの相互依存

→ 金融インフラを考えるに当たっては、提供される金融サービスとの間で相互に影響し合うことを念頭に置く必要。特に金融サービスの創出を促す金融インフラとする必要。その際、法規制、IT インフラ、インセンティブ、ガバナンス、インフラの利便性、情報流通といった観点から具体案を深堀していくことが課題。

③ ブロックチェーン技術・DLT（分散型台帳技術）・トークン化、ビッグデータ・AI などを利用したデジタル金融

→ ブロックチェーン技術等は非改竄性、トレーサビリティ、コスト競争性、透明性、ゼロトラスト、プログラマビリティなどの特性を有している。特に、ハッキング耐性の強い方法による情報管理の重要性が増す。

ブロックチェーン技術に内包されている自動執行機能を活用することによって財・サービスの提供からトークン移転による決済までを自動的に実行する仕組みは大きな可能性を持つ。スマートコントラクトは新しい金融サービスを作り出す際の一つの設計モデル。

→ ブロックチェーン技術等によって生み出される新たな金融サービスを想像／創造する能力が重要。

→ 処理技術や手段の向上を図ることによって、情報の粒度を細かくし、種類を多様化し、流通速度を速めることを通じて金融サービスの情報生産機能の高度化に結び付ける必要。

（C）新しい金融インフラを構築する際の進め方

① 標準化と統合台帳（Unified Ledger もしくは Shared Ledger）の動きへの対応

→ 標準化や統合台帳の動きが、日本やグローバル金融システムにもたらす影響を理解し、対応戦略を検討する必要。特に、グローバル・スタンダードになりそうなケースでは、最初からインナーサークルに入り、方向性の決定に関与し、コントロールしていくことが重要。

→ 国内外の利用者から選ばれる金融市場とするためにも、他国の金融インフラ等との互換性や相互運用性を確保することが必要。自国に有利に働くような仕掛けを早期の段階から組み込んでいく必要。

→ ルール・規則、プログラミング・コードなどの標準化への対応が喫緊の課題。出遅れないためにも世界で進められている標準化の動きに積極的に関与する必要。

→ 標準化の最も極端なケースとして、台帳インフラごと統合してしまうという統合台帳の考え方がある。資金決済インフラがそうした世界に移行することも想定し、誰が債務としてのマネーの発行体になるのか（中銀、民間金融機関、その他）、マネーの通貨単位の選択は発行体の国籍とリンクしている必要があるのか（リブラやステーブルコイン発行体がどの国の通貨建ても選択しうるように、中銀や民間銀行が多国通貨建て預金マネーを発行することの是非）といった論点についても、あらかじめ検討しておく必要。

② 標準化に向けた当局の積極的な関与

→ 国内の標準化を巡る動きにおいては、自主規制団体が標準化機能を十分に果たせるようにするため、当局が積極的に関与するとともに、独占禁止法適用の懸念が生じないようにするなどの工夫が必要。

（例）ISO20022 による標準化の動きへの対応。トラベルルールの採用に当たって複数のシステムが混在。

③ 情報収集・分析能力の程度による投資家の区分け

→ 保護の対象とする投資家の区分けについては情報収集・分析能力の有無によって判断する必要。デジタル空間における金融取引が増えてきた場合、デジタル空間に関する対応能力にも留意する必要。

→ 十分な情報収集・分析能力を有している場合（プロ投資家）は現状の緩やかな規制で対応。

併せて、プロ投資家の範囲の拡充・柔軟化も検討する必要。

→ 十分な情報収集・分析能力を有していない場合（一般投資家）は消費者保護を重視する観点から、規制を強化するという手法よりもむしろ、十分な情報収集・分析能力を有している金融機関等を活用することで対応。

併せて、情報開示義務の対象範囲の緩和に関する検討も必要。

→ 日本国民全体の金融リテラシーの向上も重要。

④　リテール金融とホールセール金融の区分

→　金融サービスの内容変化に伴い、リテール金融とホールセール金融の区分けについて見直しを行い、それらを区別して検討する必要。

⑤　ニーズ志向とシーズ志向

→　デジタル金融関連の技術は日進月歩。新たな金融サービスを創造する場合、「金融サービス需要者側からの求め」などのニーズ面と「金融インフラの技術的進歩」などのシーズ面の両方から検討する必要。金融サービスの需要という視点は重要だが、スマートフォンやインターネットなどの事例に見られる通り、知らないもの・現存しないものは需要側が想像できないため、新たな需要を掘り起こす観点からもシーズ志向の視点も必要。理解できないものへの関心を持つ力を社会全体が高める必要。

⑥　規模・属性などに基づく段階的な規制の活用

→　取引量などの規模、情報収集能力の差異などの属性等に基づき、規制や取扱方法に段階を設ける工夫（テーラリング・ルールなど）も必要。

（D）当面の課題・懸念事項

①　セイフティネットのあり方

→　預金には資金決済手段の提供のほか、信用創造機能の一部として機能する面もあり、他の資金決済手段とは異なる特殊性が存在。預金は強い粘着性を有していたが、デジタル化の進展に伴って低下しているとの指摘があることから、預金取扱金融機関では流動性への対応がこれまで以上に慎重に対応する必要。併せて、流動性不足などに備えた小口預金のセイフティネットについて、法規制のあり方を検討する必要。

→　預金のデジタル化に当たっては、ポイントの扱いも含む付利の状況などに対応したセイフティネットのあり方を検討する必要。

→　担保資産が必要なステーブルコイン等には信用創造機能がないのに対して、信用創造機能とセットになっているトークン型預金には別途の配慮が必要。

→　デジタル化の進展は、預金の安定性だけでなく、金融ビジネスモデルの安定性にも影響してくる。ビジネスの急拡大と同様に急激な縮小もデジタル化社会の特徴となっている。金融システム安定の観点から、こうした事態への対応を考えていくことも求められる。

②　リスクに応じたコンプライアンス対応

→　日本の金融機関では、預金などの取扱いにおいて小口・大口の区分なく、同一の基準でコンプライアンス対応しており、非効率。

→　リスクに応じたコンプライアンスを徹底する必要。場合によって、金融機関がコンプアイアン対応範囲を規定した上で、金融サービスを提

供する対応方法も一案。

③ **金融機関グループ内における情報共有の阻害要因（ファイアウォール規制など）への対応**

→ 規制緩和されたとは言え、ファイアーウォール規制などによって金融機関のグループ内の情報共有が阻害され、情報生産機能にも影響。

→ 金融分野・非金融分野の融合を見据えて、顧客情報の適切な管理体制を構築の上、ファイアーウォール規制の全廃を検討する必要。

→ デジタル金融の進展に伴い、選別すべき情報だけを取り出して連携を図ることも可能。

→ 資金決済・預金に特化したコアバンク（Narrow Bank）を分離して、コアバンクと非コア業務の間にリングフェンスを設けると情報を共有できず、金融機関の情報生産機能の強化を阻害。

④ **金融機関による DX 人材の育成・確保**

→ 金融機関は DX 推進の担い手として情報生産機能を十分に発揮するため、自らが率先して DX 人材を育成・確保する必要。

⑤ **我が国の法体系（資金決済法・金融商品取引法・銀行法等）の整理・拡充と金融制度のあり方**

→ ステーブルコインを電子決済手段として位置付けていることは評価。

→ 原油や NFT などの実物資産を裏付けとするアセット・トークンへの対応の観点から現行法を見直し、実物資産のトークン化の利用促進を図る必要。

→ 新たな金融インフラの構築を機に法体系の簡素化に努めるとともに、デジタル化に対応した金融制度（法規制・監督体制など）のあり方を検討する必要。

⑥ **暗号資産などのデジタル金融資産の投資該当判断基準の国際標準化**

→ 暗号資産などのデジタル金融資産が投資に該当するかどうかの判断基準については、国際的にみてスタンダートな基準は確立していない。

→ 我が国では将来の事業性の有無で判断。一方、米国では Howey Test 基準を採用しているものの、時代によって実際の判断にブレが生じている。

→ デジタル金融資産の投資該当判断基準については複数の段階で設定することも一案。

3．提言に当たって留意すべき事項

① **分かり易く、柔軟性に富んだ提言**

→ 扱う内容の抽象度が高い提言の場合、キャッチーなフレーズを入れるなど、分かり易さに気を付ける必要。

→　Defi などの世界は目まぐるしく変化していることから、金融制度や法規制も柔軟性を持つ必要。
→　柔軟に対応できるようプリンシプルベースの規制も一案。ただし、当局による恣意的な運用にならないような歯止めが必要。

②　全く新しい法規制・制度を構築する際の視点

→　2（C）①とも関連して、現行規制を前提に課題解決を考える方法（課題解決型）と規範的（Normative）な観点から考える方法があり、全く新しい仕組みを検討するには後者の進め方も有用。

③　短期的な視点と中長期的な視点

→　提言などをまとめる際には、(a) どのくらい先を見据えたものとするのか、(b) 短期的な視点と中長期的な視点なのか、(c) 期間にかかわらず必要となる視点なのかを考えておくことが重要。

（以上）

次世代金融インフラの構築を考える研究会

研究会メンバー（50音順）

小田（おだ）　玄紀（げんき）　氏　一般社団法人　日本暗号資産取引業協会　代表理事

廉（かど）　了（さとる）　氏　三菱 UFJ リサーチ＆コンサルティング㈱ 調査・開発本部調査部 主席研究員

小早川（こばやかわ）　周司（しゅうじ）　氏　明治大学政治経済学部 教授

増島（ますじま）　雅和（まさかず）　氏　森・濱田松本法律事務所　パートナー

山上（やまがみ）　聰（あきら）　氏　㈱NTT データ経営研究所 クロスインダストリーファイナンスコンサルティングユニットエグゼクティブコンサルタント 兼 グローバルビジネス推進センター

若園（わかぞの）　智明（ちあき）　氏　公益財団法人　日本証券経済研究所　主席研究員、理事

（事務局メンバー）

山沖（やまおき）　義和（よしかず）　氏　SBI 金融経済研究所㈱　特任研究員
信州大学　名誉教授

副島（そえじま）　豊（ゆたか）　氏　SBI 金融経済研究所㈱　研究主幹

（オブザーバー）

金　融　庁

日 本 銀 行（金融研究所）

次世代金融インフラの構築を考える研究会

開催日程

第 1 回会合（2023 年 12 月 25 日）

- 政井理事長挨拶
- 研究会メンバー紹介
- 研究会の進め方（副島）
- 事務局説明（山沖、副島）
- 論点のたたき台（山沖）

第 2 回会合（2024 年 2 月 20 日）

- メンバー報告（山上 聰 氏、小田 玄紀 氏）

第 3 回会合（2024 年 3 月 28 日）

- メンバー報告（若園 智明 氏、廉 了 氏）

第 4 回会合（2024 年 4 月 3 日）

- メンバー報告（増島 雅和 氏）
- 論点の例示（山沖）
- 討議用の論点の提示（副島）

第 5 回会合（2024 年 5 月 9 日）

- メンバー報告（小早川 周司 氏）
- 論点の例示（山沖）

第 6 回会合（2024 年 6 月 11 日）

- 報告書案の提示　（山沖）

第 7 回会合（2024 年 7 月 2 日）

- 報告書案のとりまとめ

「次世代金融インフラの構築を考える に当たっての指針」を公表して

（左上から右下へ）
山沖 義和｜ SBI 金融経済研究所株式会社 特任研究員、信州大学 名誉教授
副島 豊｜ SBI 金融経済研究所株式会社 研究主幹
若園 智明｜公益財団法人日本証券経済研究所 主席研究員、理事
廉 了｜三菱 UFJ リサーチ&コンサルティング株式会社 調査・開発本部調査部 主席研究員
小田 玄紀｜一般社団法人日本暗号資産取引業協会 代表理事
山上 聰｜ 株式会社 NTT データ経営研究所フェロー
小早川 周司｜明治大学政治経済学部 教授
増島 雅和｜森・濱田松本法律事務所 パートナー
政井 貴子｜ SBI 金融経済研究所株式会社 理事長

山沖 義和
SBI 金融経済研究所株式会社 特任研究員、信州大学 名誉教授

若園 智明
公益財団法人日本証券経済研究所 主席研究員、理事

小田 玄紀
一般社団法人日本暗号資産取引業協会 代表理事

山上 聰
株式会社 NTT データ経営研究所 フェロー

山沖　昨年 12 月に決済・金融法制・IT システムの各分野や銀行・証券・暗号資産業界で活躍されている 6 人の有識者をお迎えして「次世代金融インフラの構築を考える研究会」を立ち上げ、議論を重ねました。次世代金融インフラの将来像を描くといった大きなテーマを扱うためには、拠り所となる考え方や視点などの指針を設定するところから始めるべきという共通理解のもと、今般、「次世代金融インフラの構築を考えるに当たっての指針」をとりまとめました。
各業界等の専門家であるメンバーのご意見を集約したこともあり、指針で示した視点や留意事項は多岐にわたり、目的・視座に加えて、全部で 19 項目となっております。また、当面の課題等としても 6 項目を挙げています。
本日は、メンバーの皆さんにお集まりいただき、それぞれの専門分野の立場から指針の中で特に強調したい点、今後の研究会で取り上げてほしい論点などについてお伺いしようと思います。まずは強調したい点についてご自由にご発言ください。

若園　証券業に身を置く立場からデジタル化がもたらすメリットを挙げるとするならば、なにより従来のシステムでは不可能だった新しい金融サービスや金融商品が産まれてくるという点にあると考えます。新しい IT の力を活用して革新的なアイデアを実現することができるようになれば、金融というものが社会にいろいろと貢献ができるのではないか、本当の意味での理想形に近づいていけるのではないかと思います。
次世代の金融システムの姿を一言でいうならば、より自由で多様な発想を活かせる金融資本市場を創り出すことだと考えます。これは日本の経済成長にとって前向きな取り組みであり、国際社会でのプレゼンスを高めるためにも大変に重要なテーマだと思います。

小田　今回の報告書で特に読み手の皆様に意識していただきたいことは、ブロックチェーン技術など新しい金融技術を活用して、どのような新しい金融サービスが産まれてくるかをまずは想像してみてもらいたいという点です。それが現状の規制、例えば資金決済法や金融商品取引法の想定する世界観にフィットしないのであれば、新しい規制の枠組みを考えていこうということを報告書に示していますし、こうした視座を多くの読者に共有してほしいと考えています。

山上　自分の専門領域であるコンサルティングの立場から申し上げると、金融機関はこれまで利息収入に依存してきましたが、デジタル化の進展によりアドバイザー収入や情報仲介機能など情報生産サービスへのビジネスシフトが生じると考えています。クレジットカードの履歴を自行のための顧客モニタリングに使うというのは誰でも思い付くことですが、ある米銀は、自ら広告代理店を設立し、取引先である加盟店のために売れ筋商品の情報提供を行うというデータビジネスを始めたそうです。デリバティブについては、登場当初、投機ビジネスとみなされていましたが、その後、リスク管理手段、リスク対比で見た価

格の発見手段という金融の必須インフラとして根付きました。これも金融の情報生産機能の発露だと考えています。この事例と同様に、今後、デジタル社会における価値創造を目指した議論を深めていかねばなりません。例えば、地域型決済サービスの活用もその一つです。グローバルプラットフォーマーの掌の中でのビジネスにとどまらず、日本の競争力を高めるような舞台の仕掛けを作ることが重要です。欧州では、この点が強く意識されており、参考になると思われます。

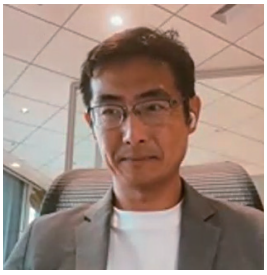
小早川　私が強調したいことは、国際化やクロスボーダー決済の進展、さらにはBISが提唱している統一台帳やシングル・プラットフォームという枠組みの検討が進むことによって、やや長い目で見て金融産業に何が生じるかを今から考察していく必要があるという点です。将来、国内外の区別なく一元化された金融インフラの運営や金融サービスの提供が可能になると、内国・外国為替という二分法的な発想が通用しなくなる可能性があります。金融インフラの利用者も当局・中央銀行も思考の枠組みやマインドセットを切り替える必要があります。ボーダーレス化を念頭に置きながら、組織やビジネスの在り方を考え直す時期に差し掛かっていると思います。

小早川 周司
明治大学政治経済学部 教授

廉　今のお話は国内の金融産業構造の見直しという論点も含んでいると思います。デジタル化の進展に伴い、銀行と商業の垣根はますます低くなり、銀行や金融の枠組み自体が変化していると感じています。銀行持ち株会社法制なども含めて法的・制度的な枠組みを再考する必要があると思っています。

廉 了
三菱UFJ リサーチ＆コンサルティング株式会社 調査・開発本部 調査部 主席研究員

増島　金融の視点だけではなく、デジタル化が進むなかで国全体としてこの大きな変化をいかに国富につなげるかが重要な論点になると考えています。デジタル化にいかに対応するかという受動的な態度を超えて、デジタル化の波をとらえてもう一段の成長を実現するという視座を産業全体が持たなければならず、それを実現するために金融はどのような機能を果たすべきかという視点で金融インフラを見直していく、そうした議論を進めていかねばなりません。産業全体のインフラという観点からは銀行や証券だけでなく、より実社会のリスクを取り扱う保険も重要な役割を果たすと考えていますので、その見直しの中には保険もしっかりと位置付けるべきです。また、デジタル化の進展によるデータを中心とした取引体系は、必然的にボーダーレス化する運命にあります。そのような市場環境のもと、魅力的な市場として日本に対して資金や資本が流れてくる仕組みをどう作っていくのか。以上は金融インフラをどうするのかということよりも一段高い次元で、日本全体の大戦略として練り上げていかなければならない課題だと考えます。



増島 雅和
森・濱田松本法律事務所 パートナー

山沖　次世代金融インフラの構築という研究会テーマにふさわしい、あるいは、それ以上の大きな課題をご指摘いただき、大変有難うございます。それでは、今後の研究会で議論していきたいテーマについてもご意見をお願いいたします。

若園　デジタルトークンがもたらす未来像です。金融資産のみならずさまざまなアセットがトークン化されることにより、従来の資産が新しい形で取引可能になります。これは、ポートフォリオの選択肢が大きく広がることを意味します。例えば、トークナイゼーションされた農産物が金融資産のトークンと並んで取引されることも可能でしょう。見方を変えると、今のシステムのもとで何が問題となっているのかを、現状の仕組みを前提とせずに自由に検討してみることが大事だと考えます。
もう一つはゲートキーパーの役割についてです。デジタル化の進展に伴い、銀行や証券会社などの伝統的な金融機関の役割は低下し、情報仲介を行う専門業者の重要性が高まってきます。それが、規制の空白地帯に生まれてくることもあるでしょう。ゲートキーパーの重要性は一層高まっていくと思います。

小田　先ほど申しましたように、暗号資産の領域については基本的には資金決済法または金融商品取引法のいずれかが適用されることとなっていますが、実際は、この２つの法律だけではカバーしきれないものもあると思います。そのため、新しい分野に対応する新しい法律を作ることも含めて、中長期的に検討していく必要があると思っています。

小早川　今回、公表した指針に記載された論点は、どれを見てもスケールが大きいほか、スコープも深く、ステークホールダーも多岐にわたります。今後、これらを深掘りする議論ができればよいと考えています。また、銀行界と証券界の間には考え方・見え方に大きな違いがあります。銀行が主導する資金決済と証券界が主導する証券決済は、相手方から見るとそれぞれブラックボックスのように感じられます。しかし、今、世界で進展している次世代の金融インフラ構築に向けたさまざまな取り組みを見ていると、日本が国際的な議論で主導権を握るためには、銀行と証券の連携が不可欠であると痛感します。

山沖　最後に、昨年６月まで日本銀行で金融研究所や FinTech センター、決済・金融システム部署で勤務され、今回も事務局メンバーとして研究会に参加した立場から、副島さんに締めくくりの発言をお願いできますでしょうか。

副島 豊
SBI 金融経済研究所株式会社 研究主幹

副島　金融システムというものは、漸進的に改善が進む、あるいは歴史を通じてよく練り込まれ完成された姿に落ち着いているというのが金融産業で働く者の常識的な見方です。ところが、通貨や金融システムの歴史を学ぶと、大きなジャンプを起こして短期間のうちに全く異なる姿に進化する、そういうことが稀に生じていることに気が付きます。江戸期から明治にかけての金融システムの大転換がその一例でしょう。金融に限らず最近の社会の変化を見ていると、実は大きな転換点を迎えているのではないか、頭を柔らかくしてさまざまな未来の可能性を考えてみる時期に来ているのではないか、これが研究会を立ち上げた動機となっています。
研究会メンバーの方々には、各分野のプロフェッショナルとして、鋭くも創造的で建設的なご意見を交わしていただき、大変感謝しております。今後、指

針に示された論点を掘り下げて議論し、次世代金融インフラの将来像を皆様といっしょに描いて行きたいと考えております。

山沖　皆様におかれましては、指針のとりまとめに当たってご協力いただき有難うございます。指針に込められた各人の思いを大切にしつつ、今年後半に再開する研究会に臨みたいと思います。

SBI 金融経済研究所 所報 vol.6

2024 年 8 月 31 日発行

発行者：SBI 金融経済研究所株式会社

住所　〒106-6013
東京都港区六本木 1-6-1
泉ガーデンタワー 13F

電話　03-6229-1001

制　作：株式会社フクイン

SBI金融経済研究所