自動認識AIの諸問題と解決へのアプローチ

東京理科大学 工学部 情報工学科 准教授 入江 豪（いりえ ごう）

■はじめに

『AI』，“Artificial Intelligence”の略称で，日本語では人工知能と訳される言葉です．簡単に言えば，人間のようにふるまうことができる機械やコンピュータのことを指しています――もはや，いまさらこんな説明をする必要もないくらい，私たちは日常的にこの言葉を目にするようになりました．

本稿をお読みいただいている皆さんもご存知の通り，この十余年の間に，AIは驚くべき技術的進展を遂げ，社会的にも大変な影響力を持つようになりました．2024年の企業時価総額の世界ランキングTop10には，GoogleやAmazon，NVIDIAなど，AI分野をけん引するビッグテック7社が顔を並べており，AIが産業の基盤として確立されていることが窺い知れます．学術界においてもその影響力は絶大で，学術会議・雑誌のインパクト指標ランキングTop10[※1]には，最高峰の学術誌として名高いNatureやScience，上位医学誌などに並んで，AI分野の学術会議が3つもランクインしています．さらに昨年，大変象徴的であったのは，ノーベル物理学賞・化学賞がAI分野の研究者に贈られたことでしょう．これらの事実からも，AIが学術界においてもその中心的分野の一つへと成長したことは疑いようがありません．もちろんAIは我々の日常生活の中にもどんどん入り込んできており，私たちは意識的にであれ無意識的にであれ，AIを使って生活する時代となりました．既にAIは人類の“パートナー”として，共生していく時代になりつつあると言っても過言ではありません．

しかし残念ながら，すべてが順風満帆というわけではありません．あまりにも急速な進化を続けてきたがゆえに――ともすれば進化を急ぎすぎたがゆえに――AIは様々な“問題”を我々の日常に引き起こしてしまってもいます．さらに，そのうちのいくつかは，私たち人類や社会，地球環境に対して，甚大な被害を及ぼす可能性さえあることが，専門家や研究機関によって相次いで指摘されてきています．

それらの問題とはどのようなものなのか．そして，AIと共生する時代に向けて，我々はそれらの問題にどう向き合い，解決していくべきなのか．様々な視点からの議論を要する難しい問いですが，本稿では技術的な観点からこの話題に触れてみたいと思います．

以降，本稿ではまず，これまでのAIの進化の歴史を概観し，現代のAIが抱えている諸問題を紹介します．続いて，こういった諸問題に対し，AIの中でも筆者の研究室で取り組んでいる自動認識技術に話題を絞り，「AIをいかにうまく“退化”させるか」によって解決を図る試みをご紹介させていただこうと思います．

■AIの進化の歴史

現代のAIが抱える問題に触れる前に，これまでAIがどのような技術的進展をたどってきたのか，その歴史を振り返ってみましょう．1956年，アメリカのダートマス大学で開かれた『ダートマス会議』にてArtificial Intelligenceという学問が誕生して以来，AIにはこれまで四度のブームがあったと言われています．各ブームがいつ始まったのか，そしていつ終わったのかについては諸説ありますが，一説によると，一度目のブームはまさにAIが誕生した1950年代後半から1960年代初頭にかけて，二度目のブームは1980年代後半から1990年代にかけて，三度目のブームは2010年前後から，四度目のブームは2022年頃から始まったとされています．一度目，二度目のブームは確実に終わりを迎えたわけですが，三度目，四度目のブームについては現在もなお続いていると考えている専門家も多くいます．実際に，AI関連分野の学術会議の参加者数や研究発表数は，第三次ブームが始まった2010年前後から右肩上がりに伸び続けており，その勢いはとどまるところを知りません．

さて，こういったブームの背景には，そのきっかけとなった出来事が存在するものです．特に重要な出来事であったといえるのは，第三次ブーム，及び，第四次ブームのきっかけとなった二つの技術革新でしょう．

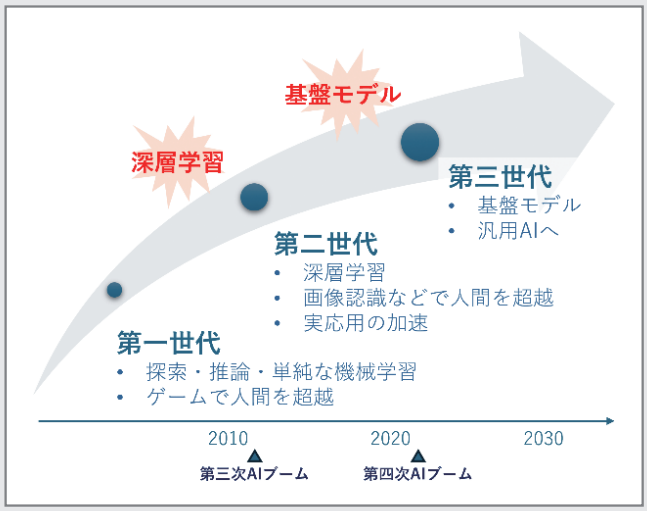
スタンフォード大学が 2021 年に発表したテクニカルレポート[1] でも，これら二つの技術革新を境界に，AI を 3 つの“世代”に分けて論じています．この二つの技術革新とは何だったのか．本稿でも AI を三つの世代に分け，各世代の特色と共に見ていきたいと思います【図 1】.

(1) 第一世代の AI：第三次ブームが始まった 2010 年代初頭までの AI．これを，先ほどの世代の話に合わせ，便宜上“第一世代の AI”と呼ぶことにします．第一世代の AI は，探索や推論，そして（現在と比較すれば）比較的単純な機械学習を核として発展してきました．

その活躍の舞台の一つは，ゲームでした．人間がその知能を活かしてプレイするものの象徴がゲームであり，ルールや勝敗が明確であったためでしょう．例えば，1997 年，IBM が開発したチェス専用のコンピュータシステム『ディープ・ブルー』が，当時のチェス世界チャンピオンに勝利したことや，2011 年には同じく IBM が開発した質問応答システム『ワトソン』がクイズ番組で人間のチャンピオンに勝利したことは，今なお多くの AI の教科書にも紹介されているほど，AI 史上に残る重要な出来事として認知されています．ゲームプレイ AI の一つの完成形ともいえるのが，恐らく皆さんの記憶にも残っているであろう『AlphaGo (アルファ碁)』です※2．囲碁は，よく知られた対戦型ボードゲームの中でも桁違いに盤面の状態数が多く，それが故に AI には攻略困難とされていたゲームでした．しかし AlphaGo は，2015 年 10 月にプロの囲碁棋士を初めて破り，続く 2016 年 3 月には，当時世界最強の囲碁棋士の一人とされていたトッププロをも破ることにも成功しました．AlphaGo が打ち出した手のいくつかは人間には相当困難なものであるそうで，プロ棋士にも模倣されて定石にもなるなど，囲碁の考え方にも変革を起こしたとも言われています．

このように，第一世代の AI はゲームにおいてはその強さをいかんなく発揮してきました．しかしながら一方で，実問題においては，我々が期待していたほどには活躍を見せられずにいたというのが正直なところでしょう．このことこそ，第二次 AI ブームがかくも短く終焉を迎えることになってしまった原因の一つであったとも言われています．

しかし 2012 年，とある事件を契機に技術革新が起こり，AI は第二世代へと生まれ変わることになります．そしてこの技術革新のおかげで，実問題でもその力を発揮することができるようになっていったのです．



【図 1】AI の進化とその世代

(2) 第二世代の AI『深層学習の台頭』：その技術革新こそが『深層学習（ディープラーニング）』であり，第三次ブームを引き起こすきっかけとなった技術です．深層学習とは，（人工）ニューラルネットワークと呼ばれる，線形関数と単純な非線形関数を組み合わせた“ノード”を，多層のネットワーク状に組み合わせて構成された複雑な合成非線形関数を，上手に学習するための技術のことを指します．基本的な理論や要素技術自体は 1980 年代から発明されていましたが，その後の計算性能やデータアクセシビリティといった情報基盤の進歩とかみ合った結果，現代になってようやくその真価を発揮することになった技術です．

深層学習を一躍有名にしたのが，画像認識技術に関する，とある国際コンペティションでした．1,000 種類の物体のうちいずれか一つが写った画像を見て，その物体の名前を解答するという画像認識タスクの精度を競うものです．そのタスクの難しさから業界の注目度も高く，2010 年に第一回が開催されて以降，毎年世界各国から腕に覚えのある研究チームが参加し，しのぎを削っていました．事件が起こったのは，その三度目の開催回にあたる 2012 年のことでした．トロント大学の研究チームが，それまで主流となっていた画像認識手法（確率モデルに基づく画像特徴表現と線形識別器によるもの）とは一線を画す，深層学習を用いた画像認識モデルを持ちこみ，二位以下に大差をつけて圧勝したのです．

この事件は瞬く間に世界を席巻し，以来，深層学習は一気に画像認識技術のスタンダードへとのし上がる

ことになります．翌年 2013 年のコンペティションではほとんどのチームが深層学習を導入し始め，画像認識や機械学習といった AI の主要分野の学会では深層学習に関する研究成果が大量に報告されるようになりました．そしてわずか 3 年後の 2015 年には，長年の夢の一つでもあった人間の認識精度をついに凌駕するに至ります．深層学習は音声音響情報処理や自然言語処理といった画像認識以外の分野へも波及し，AI 分野全域（あるいは AI 分野外にも）へと拡大していきました．さらに，当時の研究コミュニティやリーディング企業が，技術のオープン化（端的に言えばオープンソース化）を推進しており，これが定着していったことも進展を加速させる大きな要因となりました．世界中の研究者により日々先端的技術が発明され，その多くを誰もが自由に利用することができるようになったのです．

こういった流れを受けて，AI はついにゲームの枠を超え，実問題へと活躍の場を広げることになりました．情報検索や個人認証，機械翻訳，対話型エージェントなど，現在の我々の生活を支えている AI 技術には，深層学習の導入により実用レベルまで引き上げられたものが数多く存在しています．

（3）第三世代の AI『基盤モデルへ』：第二世代の AI は目覚ましい発展を遂げてきたわけですが，現在，AI はさらに第三世代へと進化し，いよいよ人間の“知能”さえも超えた“汎用 AI”に到達したと言われるほどになりました．

第三世代の AI の代表格ともいえるのが，2022 年 11 月，OpenAI によって開発・公開された汎用対話 AI『ChatGPT』です．今，「AI と聞いてイメージするものは？」と聞かれたら，これを思い浮かべる方も多いのではないでしょうか．『GPT』というのは，“Generative Pre-trained Transformer”の頭文字をとったもので，Transformer と呼ばれるタイプのニューラルネットワークを膨大な量のデータで学習した，学習済みの大規模言語モデルのことを指しています．GPT の初版が開発されたのは 2018 年で，ChatGPT としてリリースされたのはその第 3.5 版でした．その後も大幅な機能改善を続けており，例えば 2024 年 5 月に公開された『GPT-4o』では画像の自動認識機能が大幅に拡充され，画像認識専用に作られた AI を凌駕する性能を発揮しています．

この GPT の“知能”は実に目を見張るものです．2023 年時点で既に，アメリカの大学共通テストのような位置づけにあたる試験『SAT』において，名門大学への入学を見込める成績を記録したほか，米国の司法試験や医師試験でも合格できる成績を示しました．画像認識においても，単に画像の中に写る物体の名称を答えることができるだけでなく，複数枚の画像の関係を読み解き，さらにはユーモアまで理解しているかのようなふるまいを見せることが示されました．多くの労働者の業務を代替可能であることも示唆されており，OpenAI は米国労働者の 80% が少なくとも業務の 10% において影響を受けるであろうと試算しています．

GPT のように学習が済んだ状態で提供され，追加学習なく（あるいはほんの少しの追加学習を行うだけで）様々なタスクを解くことができる AI モデルは『基盤モデル（Foundation Model）』と呼ばれ，第三世代の AI の象徴となっています．現在，多くの研究団体やリーディング企業が多数の基盤モデルを開発し，さらなる性能の改善に取り組んでいるという状況です．

■いまなお残されている“破滅的”問題

AI がこれほどまでに劇的な進化を遂げた現在，「AI にはできないことはもうないのではないか」「AI は既に私たちの生活に欠かせない“パートナー”となった」と考える方もいらっしゃることでしょう．しかし，現実はそう楽観的ではありません．皆さんの中にもひょっとすると，AI に対して漠然とした不安を抱いている方も少なくないのではないでしょうか．実は，一般の人々だけでなく，多くの専門家も同様の懸念を抱いており，実際に大学や企業，研究団体が警鐘を鳴らしています．例えば，2023 年 5 月，アメリカの NPO である CAIS（Center for AI Safety）は，AI が引き起こす可能性のある“破滅的リスク”を公表しました．また，2024 年 8 月に MIT が発表したデータベースには，AI に関連する 700 以上の潜在的リスクがまとめられています．さらに，AI をリードしている企業や大手出版社も AI が引き起こす重大な社会的問題に関する報告を続々と発表しています．本稿ですべてのリスクを網羅することはできませんが，いくつかの代表的な例を見ていきましょう．

（1）脆弱性：高い精度を持つはずの画像認識 AI が，通常では考えらえれないような誤認識を引き起こすという一種の“脆弱性”があることは，比較的早期から指摘されていました．例えば，交通標識の“STOP”

サインにちょっとしたシールを貼っただけで，AI はこの標識を“制限速度 72 km”のサインであると誤認識することが示されています．このような脆弱性を持つ AI がそのまま自動運転システムに搭載されてしまったら，本来停止する必要がある危険性の高い場所で車速を上げてしまうことになりかねず，重大な事故を起こす危険性があります．また，2017 年にはアメリカのダラスに住むとある家族の子どもが，自宅の AI スピーカーに「人形のお家を買って」と話しかけた結果，実際に高額なドールハウスが注文されたという事件があり，ニュース番組に取り上げられました．真の事件はこれからで，そのニュース番組のキャスターが「AI スピーカーにドールハウスを買ってと言うなんて，かわいらしいお子さんですね」と発言したところ，番組を見ていた多くの家庭の AI スピーカーが，そのキャスターの発言（下線部）をトリガーとして一斉にドールハウスを注文してしまったのです．これはまだ笑い話で済むものかもしれませんが，場合によっては重大な問題を引き起こしうる危険な挙動です．

（2）不適切な偏り：AI が倫理的に不適切な結果を出力したという事例は少なからず報告されています．例えば，とある基盤モデルに画像に写る人物の髪色を認識させると，「金髪」の男性の画像を入力した場合にだけ，極端に「黒髪」だと誤認することが多いことが確認されています．これは「男性なら黒髪だ」という偏った先入観が，学習データに意図せず反映されてしまい，結果引き起こされてしまった倫理的問題です．

（3）情報漏洩：顔認証を始めとした生体認証は，認証にカードやパスワードなどの道具を必要としないために利便性が高く，広く実用化されている AI 技術の一つです．一方，本人の情報を直接取得する必要があるため，個人情報漏洩のリスクを伴う技術でもあります．実は，顔認証 AI に対してとある攻撃的処理を適用すると，たとえシステムに顔写真自体が保存されていなかったとしても，AI が学習した顔写真を高い写実度で復元できることが知られています．悪意ある第三者がこれを利用すれば，個人情報の不正利用につながる可能性があり，軽視できない問題です．

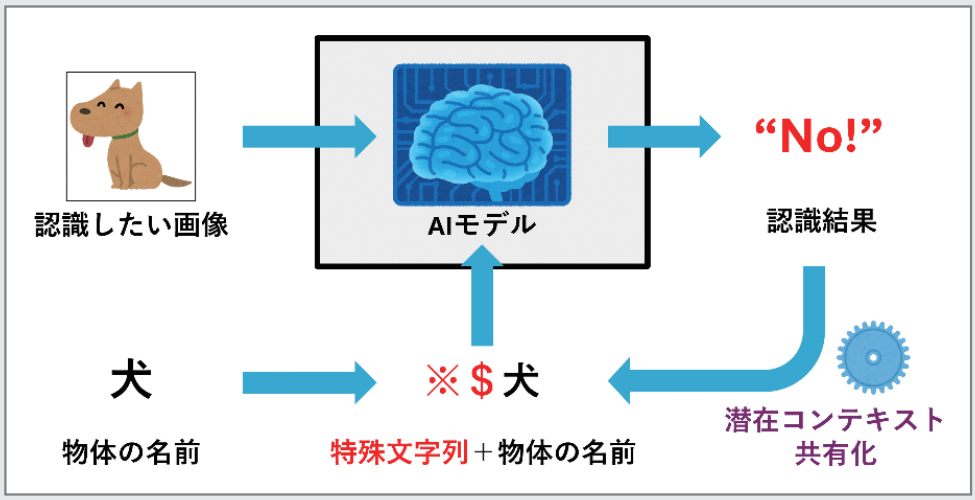
（4）エネルギー問題：現代の AI は極めて大規模であり，膨大なエネルギーを消費します．例えば，ChatGPT は Google 検索の 10 倍もの電力を消費すると言われています．2024 年 8 月の時点で ChatGPT の週間アクティブユーザー数は 2 億人を超えていますが，このユーザ数に基づくと年間消費電力は 4 億 5,300 万 kWh にも上ります．これはアメリカの 43,204 世帯分の年間電力量に匹敵し，オーストリア全体に必要な電力を 2.5 日間分丸ごと賄える量に当たるそうです．

なぜこれらの問題が解決されないまま AI は普及してしまったのでしょうか．その原因の一端は，AI の研究開発に携わる専門家でさえ，その仕組みを完全には理解できていないことにあります．技術やサービスをリリースする際には，事前にその危険性やリスクを十分検証し，対策を講じておくのが普通です．しかしながら，そもそもどういった危険性・リスクが生じうるのか，どうテストすればそれを判断できるのか，専門家にも分かっておらず，検証のしようがないのです．MIT は，顕在化したリスクのうち，リリース前に把握されていたものは全体のわずか 10% に過ぎなかったという調査結果を報告しています．つまり 90% のリスクは世に出てから発覚しているのです．さらに，AI 業界は競争が極めて激しく，“早い者勝ち”という意識が強いのも事実であるように思います．AI の進展に対しては大きな原動力となっている一方，そのリスクを広げてしまう結果を招いているのかもしれません．

■ AI を“退化”させる技術

では，これらのリスクをどのように緩和・解消すればよいのでしょう．もちろん問題は複雑で多岐にわたり，技術的な議論のみで解決できるものでもありませんが，ここでは技術的観点に限定して，一つのアプローチを論じてみたいと思います．

あえて誤解を恐れずに言えば，前述した多くのリスクは，AI が“進化しすぎた”ことによって生じたと言ってよいのではないでしょうか．例えば，エネルギー問題は AI が膨大な“知識容量”を持てるよう大規模化した結果です．また，不適切なバイアスや情報漏洩の問題も，AI がデータの微細な差異まで学習・記憶してしまうことが一因となっています．しかし，実際の応用においては，必ずしもそれほど膨大な知識量が必要ではないことも多いのではないでしょうか．例えば，自動運転システムにおける画像認識 AI は歩行者や標識を正確に認識できることが重要であり，食べ物や家具まで認識する必要はないはずです．



【図 2】不要な知識を“忘れさせる”技術[2] の概要図

そこで，私たちの研究室では，AI の性能をあえて“退化”させ，目的に合った適切な性能を持つように調整可能にする技術の開発に取り組んでいます．本稿では，その最新の研究成果を二つご紹介します．

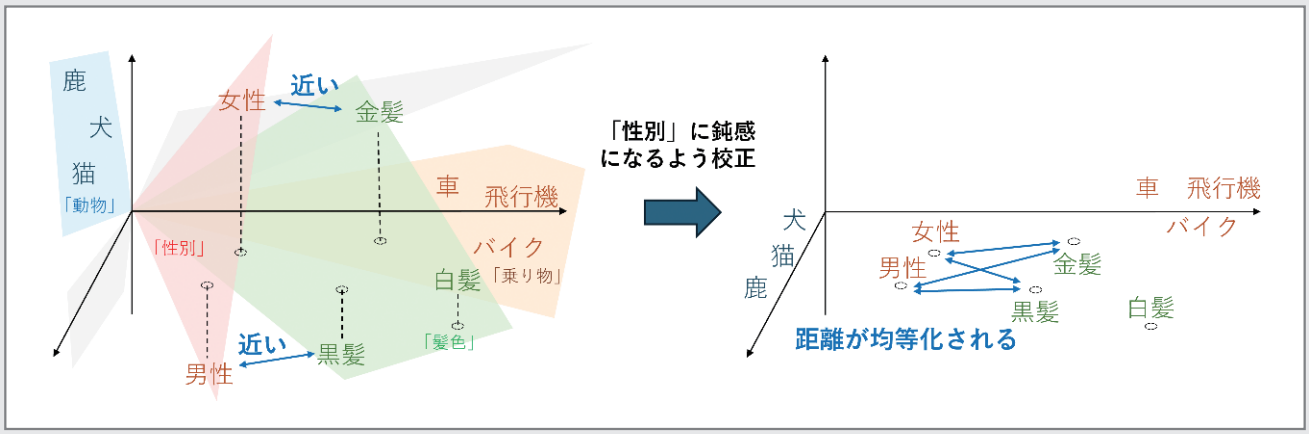
(1) 不要な知識を“忘れさせる”技術[2]：とある会社のビルで顔認証 AI を使った入館管理システムが導入されているとします．事前に社員の顔を学習しておき，ゲートに設置したカメラで通過の可否を判断する AI システムです．当然会社なので，退職や部署異動などにより，それ以上認証する必要のない人物が出てくることが想定されます．そのような人物の顔は“忘れる”よう，AI モデルを修正するのが望ましいはずです．「多くの人物を認識できることはいいことではないか」と思われるかもしれませんが，顔写真を盗み取られる危険性を考えると，認識する必要のない人物の顔データを保持し続けることはリスクになります．また，不要な知識の蓄積はモデル規模の肥大化を招き，余計なエネルギーを消費してしまうことにもなりえます．

このようなリスクを解消するために，私たちは学習済みの AI モデルから特定の知識だけを“忘れさせる”技術を開発しています．この技術は先のリスクを回避するだけでなく，「忘れられる権利」をサポートする重要な技術ともなるでしょう．

実は我々は過去に，必要な知識を順次学習しながら不要な知識を“忘れさせる”継続学習技術を世界に先駆けて実現していました[3]．しかしながらこの技術を利用するには二つの条件を満たす必要がありました：(i) 学習時に特定の“仕込み”を行えること，(ii) AI モデルの全容が既知であり，モデルのパラメータを更新できること（White-box 設定）．ところが現在主流となっている基盤モデルは既に学習が済んでいるために“仕込み”ができず，またその全容が非公開であるようなものも多く存在します．こういったモデルは先の条件を満たさないため，対処できなかったのです．

我々の最新の研究成果[2] では，この二つの条件を満たさないような AI モデルから任意の知識を“忘れさせる”新しい仕組みを考案しました【図 2】．その仕組みを簡単に説明しましょう．現代の画像認識 AI は，画像と物体名を入力として受け取り，その物体が画像内に存在するかどうかを判定します．例えば犬の画像と「犬」という物体名を与えると，AI は「含まれている」と解答するという仕組みです．今回我々が考案した方法は，物体名とともに，AI の記憶を操作できる特殊な文字列を与えます．この文字列によって忘れたいものだけを思い出せなくする――つまり，犬を忘れさせた場合，犬の画像と「犬」という物体名を与えても「含まれている」とは答えなくなる作用を引き起こすことができます．技術的なポイントはこの特殊文字列を作り出す方法にあって，我々が生み出した「潜在コンテキスト共有化」という独自の方法を使うことで，これを効率的に作り出すことができるのです．

(2)“鈍感化”して公平化する技術[4]：社会的関心を後押しとして，過去にも AI の不適切な偏りを軽減する技術は数多く開発されてきました．しかしながら従来の技術は，偏りを生じる要因が事前に特定できていないと利用できないものでした．先ほどの性差による髪色の認識の偏りの場合を例にとると，「性差によって髪色の認識に偏りが生じる」ということを知っていなければ使えない技術になっているということです．



【図 3】“鈍感化”して公平化する技術[4] の概要図

しかし当然，事前に要因を特定できるとは限りません．実際に過去問題になった例の多くは，その当時は把握していなかった要因によるものでした．

我々の研究成果[4] では，事前に偏りを生じる要因が特定できていなくても利用可能な新たな方法を考案しました【図 3】．AI の“頭の中”を見てみると，様々な概念を地図のように整理した概念空間が形成されています．意味的に近い概念は互いに近く，遠い概念は離れて配置されているような空間です．偏りの問題は，認識したい概念（例えば「黒髪」・「金髪」）と，偏りを引き起こす概念（例えば「男性」・「女性」）が，意図せず近くに配置されてしまった結果，相互に関係性がある概念だと AI が勘違いしてしまうことによって生じます．したがって解決策は，これらの概念間の距離を広げて均等化してあげることですが，事前に要因が特定できていない場合，どの概念間の距離を広げてあげればよいのかが分かりません．やみくもに概念間の距離を広げてしまうと，正常な認識能力まで損なってしまうことになります．

この問題に対して，我々は「部分空間」という幾何構造に着目した方法を考案しました．「髪色」を表す概念や「性別」を表す概念など，同じグループに属する概念の集まりは同じ部分空間，つまり概念空間の特定の一部分に集まる傾向があります．偏りの要因となる概念グループ（例えば性別）は，認識したい概念（例えば髪色）に近い部分空間に集まっているはずですので，この部分空間を特定してその影響を受けないように“鈍感”にしてあげれば，この偏りを解消することができるという発想です．部分空間は事前に与えられるわけではありませんが，スパース線形再構成という技術を使うと効率的に発見することができます．さらにこの処理は AI モデルの出力に簡単な後処理を施すだけで実現でき，モデル自体に一切変更を加える必要がないこともこの方法の大きな利点です．

■おわりに

AI の進化は，私たちの生活や社会に計り知れない恩恵をもたらしている一方で，公平性・倫理的課題や情報漏洩，エネルギー問題など，重大な問題ももたらしてきています．本稿では一つの萌芽的なアプローチを紹介しましたが，完全な解決に向けては，まだ多くの課題が残されています．AI と人間が健全に共生できる未来を築けるかどうかは，これからの私たちにかかっているといえるでしょう．本稿が皆さんの興味の一助となり，共にその未来を考えるきっかけとなれば幸いです．

参考文献

1）R. Bommasani *et al.*, “On the Opportunities and Risks of Foundation Models,” arXiv preprint, 2021.
2）Y. Kuwana, Y. Goto, T. Shibata, G. Irie, “Black-Box Forgetting,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2024.
3）T. Shibata, G. Irie, D. Ikami, Y. Mitsuzumi, “Learning with Selective Forgetting,” in Proc. International Joint Conference on Artificial Intelligence (IJCAI), 2021.
4）R. Ishizaki, S. Yamagami, Y. Goto, G. Irie, “Linear Calibration Approach to Knowledge-free Group Robust Classification,” in Proc. British Machine Vision Conference (BMVC), 2024.

※ 1　Google 社の学術会議・雑誌のランキングサイト『Scholar Metrics』における，2024 年 7 月時点でのランキングによる．

※ 2　本稿の世代区分の観点から言えば，AlphaGo は厳密には第二世代の AI に区分されるべきものですが，便宜上第一世代に位置付けて議論しています．