AIの個人の自律に関する論点とその解法

昨今のAIブームの火付け役となった深層学習の性質が孕む個人の自律への負の影響が社会に与える弊害とその解決策を紹介する。AIには他の技術と同様に課題もあるが、発展を規制などで抑制することは得策でない。利用側の認識の醸成と提供側の技術的解決の模索を同時に進めることが望ましい。

AIブームの火付け役となった深層学習手法が有する特性

深層学習手法による画像認識・音声認識の飛躍的な精度の向上、深層強化学習手法によるアルファ碁の世界チャンピオンへの勝利などにより、社会のAIに対する注目度は足元でも高水準となっている。今回のAIブームの火付け役となった深層学習は、パラメータ数・構造の点で複雑であるものの、本質的には、ある分布より抽出されたサンプルから、人間が設計したモデルを通じて、真の分布を予測することを目的とする伝統的な手法の一つである。また、実用上、多くの場合、設計者が注目する観点から仮定されたモデルを踏まえ、得られたサンプルを最もよく[1)]再現するパラメータを一意に選ぶ[2)]手法であり、学習後の出力そのものは決定論的[3)]である。以上の性質から、深層学習のみで構成されたAIは、多くの場合、学習データを構成する集団について、モデルが設定した区分に応じた統計的「一般論」を出力するため、本質的に何らかのバイアスを含むことになる。

本稿は、こうした特性を有するAI技術に関し、国内外の関係当局や民間団体がAIの倫理的課題として指摘する、自己を選択する力たる「個人の自律」への影響[4)]と、その帰結として、広く普及した精度の高いモデルが予測する人間像や関係性へと、社会が収斂していく可能性を論じ、その解決策を探る。

AIの特性が個人の自律に与える影響とその社会的影響

オハイオ州立大学の研究チームの調査[5)]によると、人間は自身がもっともらしいと考える機械出力に自己認識や自己選択をすり合わせる傾向がある。前述したAI技術の特性とこうした傾向を合わせると、技術の普及によりAIの出力に晒される機会が増し、こうした出力を素直に受け取る人間が多数を占める社会においては、精度が高いと認識され広く普及したモデルがもつ観点から予測可能、つまり、一定の視点・区分の下で一般的である、個人や関係性の割合が上昇していくことが想定される[6)]。

こうした傾向が、何ら対策を得ず常態化すると、社会は無数の閉じたコミュニティにより構成され不安定な状態となる。更に、自身の属する閉鎖的コミュニティは自己で選択したものですらない。現在、インターネットの普及に伴い、無数の閉鎖的コミュニティの乱立に伴う意見の純化・先鋭化の傾向が指摘されている。AI技術が普及した社会においては、これと同様の結果が実生活においても生じる可能性がある。AI社会のリスクシナリオの場合、自身で意見を選ぶ過程を伴わない点でネット上よりは先鋭化の傾向は穏やかなものとなることが想定される。しかし、社会的関わりがない対象に個々人が関心や共感を持つことは難しく、例えば選挙などで冷徹な判断を下し易くなることが考えられる。そうした分断された社会では、社会の安定性を過度に中央に委ねることとなりかねず、その影響を過小評価はできない。

なおAI社会では、人種・性別などによる従来からの差別や偏見にとどまらず新しい閉塞感に繋がる可能性を含むことを強調したい。想起し易い例としては、AIによる精度の高いモデルで、特定の嗜好品を購入する頻度の高い人は窃盗を犯すリスクが高いとされた場合、それが当人にとっては犯罪の抑止力になっているケースでも、社会的にこうした行為が抑圧されることが考えられる。また、差別を助長することへの防御策としての、複

NOTE

1) 最も高い確率で出力する、モデルとの何らかの意味での平均誤差が最も小さい、など。なお、深層学習などの場合、モデルの複雑さゆえ厳密にこうした点を得ることはない。
2) こうしたパラメータの候補の平均を取る手法などもあるが、どちらにしろ出力パラメータを一意に定めることに違いはない。
3) 後に紹介するように出力が確率的な手法もある。この場合もバイアスは含むが、決定論的な手法に比べるとその傾向は緩和される。
4) 例えば、総務省「国際的な議論のためのAI開発ガイドライン案」4.AI開発原則⑦。
5) https://hbr.org/2016/04/targeted-ads-dont-just-make-you-more-likely-to-buy-they-can-change-how-you-think-about-yourself
6) 例えば、法学者の山本龍彦氏は、「AI社会では、我々の個人的なアイデンティティがある特定の「集団」の一般的傾向により逆規定されるようになる」としている（「ロボット・AIと法」第4章）。
7) 無論、その時点で認識できる倫理的に問題のある機構として検知するためには有効。
8) T. Osogami and R. Raymond, "Determinantal reinforcement learning," in Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, Hawaii, USA January 2019 (to appear).
9) Personal Data Storeの略。個人データを自らの意志で蓄積・管理するための仕組み。
10) 学習データが多様でも前述のとおり出力にバイアスを含むが、データそのものにバイアスがあるよりは多様な出力となる。
11) 例えば、「人工知能と人間社会に関する懇談会」報告書（案）p.14 注釈10。

雑な出力に透明性を持たせる技術[7]は、その技術が何らかの因果関係を明示することなく、単に出力結果に関する特徴を明らかにするにとどまるならば、かえってその特徴を持つ人が疎外され、この新しい閉塞感をむしろ助長しかねないことに配慮が必要である。

AIによる弊害を軽減する可能性のある技術的手法

こうしたAI技術の負の側面を軽減する手法としては、広く社会に技術的特性を周知することは言うまでもないが、多様な応答を意図して出力する機構の設計や、少なくともデータの多様性は確保するための社会システムの構築などがまず考えうる。

前者として具体的には、深層強化学習手法として、関わるエージェントのとる行動が異なるほど価値が高いとして設計したモデルがある[8]。本モデルは、同時刻でのエージェントの戦略だけでなく、時系列の出力を確率的とする点でも、出力が決定論的であるものに比べ、多様な出力を返すことができる。

後者としては、多様な人がデータを提供するインセンティブを与える施策が求められよう。その際、他者保有のものを含め個人に関する情報は本人に帰属するとの考えを前提に設計されたPDS[9]や情報銀行を広く普及させた上で、そこで情報がやりとりされる際の価値指標として、情報量を用いることが考えられる。情報量とは、ある情報を得ることで不確実であった知識がどの程度確実となるかを測る指標である。具体的には確率の逆数に対数を取った量であり、得た情報の出現頻度が小さいほど大きくなる。特に情報銀行は、他者が個人に帰属する情報を使用する場合の便益（報酬）を本人に返す仕組みであるが、この便益を、貸し出す情報により個人が特定されるリスクに応じ定めることで、少数派に対し、こうしたサービスを使用する、もしくは、事業者に多くの情報を提供するインセンティブを与え、AI技術の背後にあるデータの多様性を担保することができると考えられる[10]。また、広くこうした方式を採用することで、AI技術の課題を社会へ周知することにも繋がると考えられる。

技術の課題は出来る限り技術により解決すべき

本稿では、AI技術の進展に伴い、社会全体がある視点における一般的な傾向へ帰趨していくリスクを取り上げた。こうしたリスクの起点である個人の自律への課題は、例えば広告による影響[11]など従来から存在するものだが、AI技術の精度や認知度の点で、より自己決定に与える影響が強いことが想定され過小評価はできない。

しかし、AI技術をただ排除するということでは進展はない。特に生産年齢人口の減少など課題先進国である日本は、人手不足の解決に貢献するといわれるAI技術から大きな恩恵を享受できると考えられ、技術発展による革新を排除すれば損失は大きい。AI技術による課題に対し、安易に規制などで発展を阻害すべきではなく、現行の技術による恩恵を享受しながら、利用する側の理解の醸成を図りつつ、提供する側が出来る限り技術により解決し、課題をも発展に昇華することが求められる。

Writer's Profile

中野 留里 Ruri Nakano
金融イノベーション研究部
副主任研究員
専門は中央銀行業務、AIに関連した研究・開発
focus@nri.co.jp