人工知能研究者として私たちがすべきこと

丸山　宏

1.技術悲観論

英 Economist 誌のクリスマス特集号は、”Pessimism v progress” というタイトルの技術悲観論[1]から始まります。新しい技術は私たちの社会をよりよくしていくはずだったのに、顔認識技術によってプライバシーが侵害され、フェイクニュースによって民主主義の根幹が脅かされ、Uber や Amazon のビジネス最適化によって労働者の労働環境が悪化し、貧富の格差が増大し、新たな管理国家が生まれようとしているのです。

産業革命時代のラッダイト運動など、技術革新に悲観論はつきものです。しかし「新しい技術には必ず陽の部分と陰の部分があり、技術そのものには責任はない」と突き放してはなりません。現状の情報技術、特に機械学習や最適化については、曖昧な報道や誤解によって過度な期待や行き過ぎた悲観論があり、その結果、技術のもたらす真の脅威が見えにくくなっています。私たちは人工知能研究に携わるものとして、健全な議論を推進する責務があります。IEEE の倫理指針[2]や、人工知能学会の倫理指針[3]は、研究者・技術者が持つべき普遍的な倫理指針を示していますが、上のような現状を考えると、喫緊の課題として以下の 3 点を、人工知能研究者に皆様に心がけていただきたいと思います。

(1) 技術を正しく伝えること
(2) 正しい理解に基づく現実の脅威について考察を深めること（適切に怖がる）
(3) 人工知能研究が明らかにする新たな問いに誠実に取り組むこと（見たくないものを、見る）

2.正しく伝える

2019 年 4 月の IEEE Spectrum 誌には、”How IBM Watson Overpromised and Underdelivered on AI Health Care”という記事[4]が掲載されています。IBM Watson が顧客に過大な期待を抱かせ、その結果プロジェクトが相次ぐ失敗に終わった、というものです。そこで聞かれた多くの批判が、「Watson は真の人工知能でない (the product isn’t “real AI”)」というものだったそうです。

残念なことですが、「人工知能」という言葉の曖昧さが社会に大きな混乱を招いているように思えます。人工知能という学術領域が存在することは疑問の余地がありません。人工知能研究に携わる我々研究者にとって、「人工知能」とは知性を模倣する機械を作ることによって知性を理解しようという研究領域です。

しかし、物理学の応用製品を「物理学」と呼ばないのと異なり、人工知能研究から派生した技術の応用製品（自動運転車、自動翻訳機など）を「人工知能（AI）」と呼ぶことがあります。ここに「AI」という言葉の悲劇があります。「AI」が発話者の意図によって様々な意味で使われるからです。研究者にとって「AI」とは研究領域であり、それ以上でもそれ以下でもありません。しかし、AI 派生技術に基づく製品・サービスを提供する企業には、それら製品・サービスを「AI」と呼ぶ明確な動機が存在します。「AI エアコン」や「AI 電子レンジ」などは、そうでない製品・サービスよりもあたかも優れたものという印象を（少なくとも第 3 次ブームが高まっている現在では）与えるからです。これに加えて、「AI」はまた、まだ見ぬ想像上の機械、汎用人工知能を指すのにも使われます。シンギュラリティや「AI に支配される世界」などの議論はこの「AI」の第 3 の意味に基づくものです。

人工知能研究に携わる者として私たちには、このような曖昧な言葉による混乱を最小にするために努力する責務があります。研究者が機械学習技術（あるいは最適化技術）を指して「AI」と言ったのに、顧客（あるいは社会）がそれを汎用人工知能と解釈したとしたら、それは誰の責任でしょうか。私たちは、現在の人工知能研究がもたらす個別の技術を指す言葉、たとえば「機械学習」や「最適化」を使い、それらの可能性と限界を正しく伝えるべきではないでしょうか。

よく聞く言い訳が「顧客（視聴者）は『機械学習』と言っても理解してくれない」というものです。しかし、機械学習を「AI」と言い換えたら、理解してもらったことになるのでしょうか。今は人工知能ブームであり、毎日の報道で「AI」という言葉を聞かない日はありません。そのため、多くの人は「AI」と言われるとなんとなくわかったような気になってしまいます。聞き手は「AI」という言葉を自分のイメージで勝手に解釈します。その誤解を、私たち人工知能研究者は自分たちに都合よく使っていないでしょうか。今の人工知能ブームを、製品・サービスを売るための方便や、競争的資金を獲得するための機会として捉えることがあったとするならば、それは厳に慎むべきだと思います。

3.適切に怖がる

技術を正しく伝えないと、ピント外れの脅威論が独り歩きします。最近はあまり聞かなくなりましたが、初期の人工知能脅威論にはいわゆる「トロッコ問題」を扱った議論が多くありました。トロッコが直進すると 1 人の人を轢いてしまい、ポイントを切り替えると 3 人の人を轢いてしまいます。ポイントを切り替えるべきかの判断を機械に任せてよいのか、という議論です。話としては面白いですが、そのような特殊な状況の議論することで、本来すべき議論の方向性が逸らされてしまい、自動運転車の安全性を向上しない、という議論[5]があります。そもそもそういう状況に陥らないように事前に手を打つのが正しい工学的センスでしょう。

「AI で仕事が奪われる」という議論もよく聞きますが、これは AI に限った話ではなく、自動化による労働の変化を指しています。「自動」がつく機械、すなわち自動販売機、自動改札機、自動洗濯機、自動ドアなどは、多くの仕事を過去のものにしました。これからも人間がやるべきでない重労働、危険な労働、単調な労働などはどんどん機械に置き換わっていくでしょう。しかし、これらはあくまでも道具としての機械をプログラミングし作業の一部を代替させるという話です。人間と同様に教えられた仕事を覚え、自ら考えて新たな事象に対応するするロボットを想定しているとすれば、それは SF の世界であり現時点での現実の脅威とはいえません。

一方で、現実の技術である SNS やそれに伴う機械学習・最適化技術は大きな脅威を社会にもたらしつつあります。真の脅威はどこにあるのでしょうか。

人工知能は人間の知性を理解しようという試みの 1 つです。未だに人間の心の働き、特に自我意識や記憶などの仕組みの多くはよくわかっていませんが、人工知能や認知科学の研究によって、人間の心の弱さについては多くのことが知られるようになりました。人には認知バイアスがあり、この認知バイアスを利用することで、人々の意思決定に影響を与えることができます。この技術は、広告の世界で急速に進歩していて、特にデータの得やすい Web 上で、人々を特定のサイトに誘導することは日常的に行われています。

歴史家であり元数学者でもある京都大学名誉教授の林晋先生は、SNS やその裏にある機械学習・最適化の技術が、核兵器に匹敵するような、人類社会に対する大きな脅威になりつつある、とおっしゃいます。ユヴァル・ハラリの「ホモデウス」[6]は、人類社会が SNS の発達により、人間の尊厳や自由意志を根本的価値とする「ヒューマニズム」の時代から、「データイズム」の世界、すなわち自分が何が欲しいか、自分が何をしたいかを機械の予測に任せることによって気持ちよく暮らせる世界に移行しつつあると主張しています。

しかし、「データイズム」の世界には大きな危険があります。ハラリが指摘するように、自分の情報を機械に開示すればするほど、機械はより正確に人々の反応を予測することができるようになり、このような予測を使えば、人々の判断を誘導することも可能になります。私たちの判断がより機械に依存するようになれば、それを利用して社会を支配しようとする人が現れるでしょう。ヒットラーやスターリンが、SNS を利用できたとしたら、何が起きたでしょうか。すべての人の需要を正確に予測しコントロールできれば、社会主義の計画経済が破綻することはなかったでしょう。中国は「情報技術で統制された社会主義」を民主主義よりすぐれた統治形態として証明しようとしているのではないでしょうか。

どちらの統治形態がより平和で繁栄した社会をもたらすのか、には議論の余地があるでしょう。民主主義は決して無謬ではなく、むしろ欠陥だらけの統治形態だからです。しかし、民主主義を失うことが何をもたらすかは、私たちは 2 度の世界大戦という大きな代償を払って学んだはずです。民主国家どうしが戦争になることはない、という民主的平和理論（Democratic Peace Theory）には一定の説得力があります。もし、平和を守るために我々が民主国家でありつづけることが重要であるならば、我々の自由意志に介入するような技術を野放しにしてはなりません。我々は人工知能研究者として、これらの点について警鐘を鳴らし続ける必要があると思います。

4.見たくないものを、見る

New York Times の 12 月 5 日の記事[7]には、“Our brains are no match for our technology”というものがあります。生物学的な仕組みからなる人類の知性は、非常にゆっくり進化します。21 世紀に生きる我々の思考は、数万年前に必要だった感情や直観に、いまだに強く依存しています。生身の人間に比して事実上無限の記憶・計算能力を持つ情報システムが、多くの面で人間の思考能力を上回ることは受け入れなければなりません。したがって我々は、自分たちの心の弱さをよく理解し、機械がそれにつけこんで私たちの思考に介入しようとしたときには、それを正しく検出できなければなりません。幸いなことに、機械学習や最適化の技術が進歩したことによって、我々の心のどこに弱さがあるのか、あるいは我々の議論のどこに曖昧さがあるのかをより明確に語ることができるようになってきています。

先日、10 月 10 日に機械学習関連の 3 研究会が、機械学習と公平性に関する声明[8]を発表しました。声明のポイントの 1 つは、公平性を機械学習の問題の 1 つとして捉えることによって、公平性の概念をより明確にすることができる、というものです。フェアであるということはどういうことでしょうか。私たちは「それは不公平だ」と不満を述べることがあ

りますが、公平であるとは何なのか、真剣に考えたことがあるでしょうか。例えば、会社の採用や昇進において女性が差別されている、というのはどういう場合に言えるでしょうか。採用の際に性別を考慮せずに判断すればよいという考え方もあります（プロセスの公平性）。一方で、採用の結果として応募者の女性比率と、採用者の女性比率に統計的な有意差があったらなんらかの差別があったとみなす考え方もあります（結果の公平性）。情報技術、特に機械学習の技術を社会に導入するにつれ、私たちは今まで定義を曖昧にしていた概念について、真正面から取り組まなければなりません。

深層学習ではその計算結果がなぜその値になったのか、説明するのが困難だと言われます。しかし、説明とは何でしょうか？ 現在の深層学習のほとんどはデジタル計算機の上で計算が実行されるので、同じ入力、同じ訓練データセット、同じハイパーパラメタ、それに同じ乱数初期値を与えて計算させれば必ず同じ結果が得られます。ですから、入力から出力に至る計算過程をビット単位で再現することは原理的に可能です。Gilpin ら[9]によれば説明とは解釈可能(interpretable)でかつ完全(complete)という、2 つの矛盾する要求を満たしたものです。解釈可能性は、受け手の能力や背景知識に依存します。専門家にとって解釈可能な説明が、一般市民にとっても解釈可能とは限らないからです。従って、説明、ひいては解釈可能性の問題は、私たちの認知能力はどれだけか、という問題に（少なくとも部分的には）帰着されます。自分たちの能力の限界を明らかにする、という努力なしには、説明の問題は語れないのです。

同様の議論が、最適化の文脈においても可能です。強化学習によって自動車の自動運転を行ったとしましょう。この効用関数は「安全に目的地へ到達する」というものになるでしょうが、その際安全を最優先するとどうなるでしょうか。まったく動かない車になってしまいます。従って、車が動くという効用と、安全性とのトレードオフを明示的かつ定量的に指定しなければなりません。今の私たちの社会は、安全でかつ動く車、というものが、あたかも両立するような幻想を持っています。最適化という言葉で技術を捉えれば、私たちは様々なトレードオフについて何等かの意思決定をしなければならない、ということに気が付きます。

このように、機械学習や最適化技術の発展によって、今まで曖昧にされていた社会の価値観の議論の重要性がクローズアップされてきています。このような議論は、往々にして深い考察を必要とし、面倒なものになりがちです。私たちは、そのような新たな問いから目をそらさずに、誠実に取り組まなければなりません。

5.まとめ

私たち人工知能研究者が喫緊にやらねばならないことは、以下の 3 つだと思います。

(1) 技術を正しく伝えること
(2) 正しい理解に基づく現実の脅威について考察を深めること（適切に怖がる）
(3) 人工知能研究が明らかにする新たな問いに誠実に取り組むこと（見たくないものを、見る）

人工知能研究者の皆様の、社会に対する責任の議論がより進むことを期待します。

参考文献

1. “Pessimism v progress,” *Economist*, Dec 18th 2019 edition.
2. “Ethically Aligned Design — A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition,” IEEE, 2019.
3.「人工知能学会 倫理指針」について，人工知能学会.
4. Strickland, Eliza. “IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care,” *IEEE Spectrum*, 2019, 56.4: 24-31.
5. Hong, Jason, “Is the Trolley Problem Useful for Studying Autonomous Vehicles?” *Comm. ACM*, May, 2019.
6. Yuval Harari, *Homo Deus: A Brief History of Tomorrow*, 2018.
7. Tristan Harris, “Our Brains Are No Match for Our Technology,” *The New York Times*, Dec. 5, 2019.
8. 人工知能学会 倫理委員会、日本ソフトウェア科学会 機械学習工学研究会、電子情報通信学会 情報論的学習理論と機械学習研究会, “機械学習と公平性に関する声明,” 12/10, 2019.
9. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L.: “Explaining Explanations: An Overview of Interpretability of Machine Learning,” *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2018.

出典：CNET Japan ブログ

丸山　宏　プロフィール

所属

株式会社 Preferred Networks PFN フェロー

Shizuoka University Visiting Professor

工学博士(1995 年)

1983 年　東京工業大学修士課程修了
同年　　日本アイ・ビー・エム入社
　　ジャパン・サイエンス・インスティテュート（後の東京基礎研究所）にて，人工知能，自然言語処理などの研究に従事
1997-2000 年　東京工業大学　情報理工学研究科　客員助教授　XML，Web サービス，及びセキュリティの研究・開発・標準化を行なう
2003-2004 年　IBM ビジネスコンサルティングサービス株式会社へ出向
　2006-2009 年　東京基礎研究所所長。執行役員
　2009-2010 年　キヤノン株式会社　デジタルプラットフォーム開発本部　副本部長
　2011-2016 年　大学共同利用機関法人　情報・システム研究機構　統計数理研究所　教授
　2016-2018 年　株式会社 Preferred Networks　最高戦略責任者
　2018 年 4 月-現在　株式会社 Preferred Networks PFN Fellow

『新　企業の研究者をめざす皆さんへ』
　丸山宏（著）　近代科学社