レクチャーシリーズ「シンギュラリティとAI」にあたって

山川 宏
（ドワンゴ／全脳アーキテクチャ・イニシアティブ）

当学会誌において，技術的特異点（テクノロジカルシンギュラリティ：以下，シンギュラリティと呼ぶ）というテーマを初めて取り上げたのは，2013年5月号の特別企画「シンギュラリティの時代：人を超えゆく知性とともに」であった．人工知能がもたらすシンギュラリティとは，人工知能の急速な発展の帰結として，従来の傾向に基づいての技術進歩の予測が通用しなくなる事象である．

その後のわずか4年間の間に，人工知能が社会に与える不便益を最小化しつつその便益を最大化するための議論が国内外を問わず大きく進展した．そこでは，人工知能の専門家だけでなく，経済学，法学，政治学，哲学などのさまざまな分野の専門家との連携により議論が進められている．今後より高度な人工知能を実現することで人間の認知能力では対応策を見つけ難いグローバルな課題においてすら持続可能な解決策を見いだすことなどにも役立ち得るだろう．しかし現実的にはまず，例えば以下のようなリスクに対する議論が進んでいる．

- 多くの場合に人工知能の導入に利益があるにもかかわらず，事故などの例外的状況の扱いが困難なために導入が難しい自動運転などの課題
- 人間が人工知能に対して過度に感情移入を行い，コミュニケーションをするのは人工知能だけでよいというような，“人工知能依存”が起きる可能性
- 既存の職業が早いペースで人工知能に置き換えられることに対応して，新たに生まれる業務に対しキャッチアップするスキル獲得が困難になる問題

こうした現在の議論は，すでに実現されているおおむね道具的な人工知能およびその延長線上の技術を起点として行われることが多い．もちろんそうした議論を進めることは有益かつ急務ではあるが，人工知能がもたらす長期的ではあるが甚大な影響（ある意味でこうした影響自体をシンギュラリティと呼んでもよいかもしれない）への注意が逸らされる危険性をはらんでいる．

人工知能技術の長期的な進展によりどのような影響があるかは簡単には語ることはできず，むしろそれ自体を本特集の著者の皆様に語っていただくことを期待している．しかしながら，思い当たる例をいくつかあげるとすれば，例えば人工知能の汎用性が高まることで人間の職業的な価値が完全に喪失したり，高度に自律的な人工知能に対して人による十分な制御が困難になったり，検証不能な高度に複雑化した人工知能システムがサイバー攻撃に対してぜい弱になったりする可能性もある．

こうした長期的な影響が，確実に100年以上先と想定されるなら，いったん棚上げしておいても問題はない．しかし現状の人工知能技術の進展は極めて急速で指数関数的ですらある．そのためその時期を安易に遠いものとは想定できない．技術的に見ると，長年にわたって記号的な人工知能が突き当たっていた子供のような知能を実現できないというモラベックのパラドックスが存在した．しかし，深層学習によってこの障壁を突破できたことで，今や人工知能が人間レベルの知能を包括的に凌駕するまでの距離感が小さくなっている．さらに人工知能はビジネスにおいて実用的なレベルとなり，その性能改善によって経済的な価値を生み出すことで再投資のサイクルが加速している．こうしたさまざまな加速要因を考慮すれば，現段階において長期的な影響を安易に検討のスコープ外とすることはできない．

人工知能のもたらす影響全般についての議論は2017年現在において世界的に高まってきており，いくつかの議論ではすでに長期的な影響がスコープに入ってきている．例えば本年の1月に米国のFuture of Life Instituteで主催されたBeneficial AIにおいて提案されたアシロマAI原則においてもその最後の5条において以下のような「長期的な課題」が示されている．

19）能力に対する警戒：コンセンサスが存在しない以上，将来の人工知能がもち得る能力の上限について強い仮定をおくことは避けるべきである．

20）重要性：高度な人工知能は，地球上の生命の歴史に重大な変化をもたらす可能性があるため，相応の配慮や資源によって計画され，管理されるべきである．

21）リスク：人工知能システムによって人類を壊滅もしくは絶滅させ得るリスクに対しては，それぞれの影響の程度に応じたリスク緩和の努力を計画的に行う必要がある．

22）再帰的に自己改善する人工知能：再帰的に自己改善もしくは自己複製を行える人工知能システムは，進歩や増殖が急進し得るため，安全管理を厳格化すべきである．

23）公益：広く共有される倫理的理想のため，および，特定の組織ではなく全人類の利益のために超知能は

開発されるべきである．

関連して，本誌次号（2017 年 9 月号）においては，人工知能とロボットによる社会への長期かつ広範な影響を，技術開発側と人文社会科学側の両面からの議論を含む「AI 社会論」特集が企画されており，この中では，アシロマ人工知能原則に影響を与えた Stuart Russell 氏らによる「ロバストで有益な人工知能のための優先的な研究事項」の和訳も掲載される予定である．

さらに本学会倫理委員会においても本年 2 月公開の人工知能学会倫理指針の中で「9（人工知能への倫理遵守の要請）人工知能が社会の構成員またはそれに準じるものとなるためには，上に定めた人工知能学会員と同等に倫理指針を遵守できなければならない.」というように，人工知能の長期的な発展を踏まえた言明を行っている．

現在の急速な技術進展の中に身を置くと，すでに「シンギュラリティ」という言葉すらやや古臭く感じてしまうが，本特集ではあえて「シンギュラリティと AI」というタイトルを採用した．なぜならこうした時期に国内随一の人工知能についての専門家集団である本学会から，人工知能の中長期的な社会への影響を考えるうえで参考となる技術的な見解を提供することは，社会的に見ても意義は高く，そうした部分に焦点を合わせたいと考えたからである．つまり本レクチャーシリーズにおける「シンギュラリティ」とは，今後の中長期的な人工知能の技術的発展がもたらす，社会への影響の全体を指しているのである．

こうして技術進展が加速している現状においては，専門家であっても技術動向予測は非常に難しい．例えばコンピュータ囲碁がトッププロを凌駕するまでに 10 年以上を要すると見る専門家は多かった．またシンギュラリティの定義からして人工知能の専門家内においても意見は分かれているし，そもそもシンギュラリティはあり得ないとか，定義し得ないという立場もあるだろう．

以上を踏まえれば大変に難しい注文ではあるが，執筆者の皆様には人工知能の専門家として個人的な思想の表明をしつつも，可能な範囲で技術的な裏付けや考察を含んだ形での議論展開いただき，本学会として比較的共有し得る見解を探りながらの情報発信をお願いした．そして特に，シンギュラリティに至るまでに克服せざるを得ない，もしくは克服し得ない，技術的課題（技術障壁）を特定し，その課題の克服に向けたこれまでの研究動向と，それが最速でどの程度で実現し得るかなどについて，できればスケジュール感を伴った形で表明いただくようお願いした．その他にも，触れていただけると望ましい項目として以下のようなことがあることをお伝えした．

- ●技術的に扱い得るシンギュラリティの定義とは何か，そもそも定義し得るのか？
  - ○何をもってシンギュラリティが起こったといえるのか
  - ○人工知能が，包括的な意味で人間の知能を超えるとは何か？
- ●シンギュラリティへの到達に関わる技術的要因
  - ○克服せざるを得ない技術的課題（技術障壁）
  - ○ある時期までもしくは永遠に起きない・もしくは起きる技術的な根拠
  - ○ステージゲート：何が実現されれば，カウントダウンが進んだといえるか
  - ○シンギュラリティに向かう技術ロードマップ（必ずしも年は特定せず）
- ●シンギュラリティがもたらすインパクトは何か？
  - ○社会へのインパクトが大きい技術要素は何か（汎用性，自律性，言語意味理解など）
  - ○そうしたインパクトはシンギュラリティ以前からの影響と連続的か
  - ○甚大なインパクトは何か（人類絶滅リスクなど）
- ●望ましいシンギュラリティとは何か？
  - ○誰がシンギュラリティを起こすことが望ましいのか
  - ○逆に，望ましくないシンギュラリティのシナリオは何か
- ●人工知能はどのように開発されるべきか？
  - ○リスクを最小化し利益を最大化する人工知能開発とは
  - ○AI 自体の開発原則・AI 開発者のガイドライン
  - ○安全性のための人工知能技術（検証可能性，説明や理解の可能性，制御可能性，バリューアラインメント，など）
- ●シンギュラリティに関わるさまざまなバイアス
  - ○立場（人工知能専門家，マスメディア，その他）上議論し難いテーマの存在
  - ○認知バイアス（正常性バイアスなど）
- ●人工知能研究開発において，今後数年以内に起こり得る大きな出来事は何か？
- ●そこで改めて，シンギュラリティはいつ起こるのか？
  - ○最速で，遅くとも，最もありそうなのは，起こらない？

いずれにしても，今後の人工知能の急速な発展は中長期的に人類社会に対してさまざまな大きな影響を与えるだろう．もし仮にいわゆるシンギュラリティが明日起きたとすれば世界は大混乱になるだろう．しかし逆に，それが数十年以上先に起こるのであれば，比較的ゆっくりとした変化に社会や人々は順応していけるだろう．だが，その場合には，今後の人工知能研究に停滞が予想されるともいえるので，専門家の視点から冷静に技術的な難しさを指摘して，過剰な期待を抑制しつつ現実的な応用価値を示していくべきである．

本レクチャーシリーズを通じて，読者の皆様には，自分達が何らかの形で関わる人工知能の技術発展の先に生じ得るさまざまな中長期的な社会的影響にも目を向けていただく機会となれば幸いである．