特　集「AI and Society」

長期的視点で考える社会へのインパクト

Safety and Benefits in the Long Term

松井　拓海　東京大学教養学部
Takumi Matsui　College of Arts and Sciences, The University of Tokyo.
matsutakubb@gmail.com

Keywords: artificial intelligence, AI safety, industrial revolution, human and AI partnership, beneficial AI.

1. は　じ　め　に

AI and Society 2日目の特別セッション「人工知能への社会へのインパクト」のセッション2「長期的視点で考えるべき社会へのインパクト」では，ケンブリッジ大学・CSERエグゼクティブディレクターのショーン・オヘガティ氏の司会のもと，Google DeepMind・FLI共同創設者のヴィクトリア・クラコフナ氏と東京大学教授・理化学研究所革新知能統合研究センター社会における人工知能研究グループ長の中川裕志氏，IBMトーマス・ワトソン研究所上級研究員・パドヴァ大学教授のフランチェスカ・ロッシ氏のパネリスト3名から15分ずつの講演があった．その後，総合討議が行われた．本稿は講演内容を要約して報告する．

2. AIの安全性のために

クラコフナ：AIの安全性とは，私達がAIに行ってほしいことを，AIがそのとおりに行ってくれることです．AIの安全性に関する長期的な問題は，どうやって常識を具体化していくか，人間によって共有されている価値観をどうやってAIに共有していくか，誰の価値観をAIのシステムと一貫させればよいのか，もしくはすべての人との価値観と一貫させることができるのか，AIに何を理解してほしいのかということです．

将来的な問題を回避し，AIのベネフィットを享受するためには多くの人々の価値観をすり合わせなければなりませんが，最適解が必ず望ましい結果になるとは限りません．また解決には時間がかかるため，今からこの問題に取り掛からなければなりません．

長期的なAIの安全性に関しての一つの重要な研究分野はスペシフィケーションです．スペシフィケーションとは，人の複雑な嗜好をAIシステムの中に明確化させていくことです．システムの報酬機能が十分に規定されていないと，人間が意図した行動をAIがとらない危険性があります．例えば，報酬を得るというインセンティブがなくてもシャットダウンできるAIが理想的です．物を運ぶときは壁を壊さないように，人に当たらないように，といった制約のすべてを特定するのは難しいです．そこでAIには何らかのヒューリスティックスをもって環境を破壊しないように設計しなければいけなくなります．

もう一つ研究分野としてはロバストネス（堅牢性）があります．人間の嗜好を明確にしたうえで，どうやってAIは人の嗜好を理解することができるでしょうか．例えば，学習時と実際の実用場面ではデータは必ずしも同じ形では出てこないという問題があります．また，AIには自ら探索して学習してほしい一方で，実際のタスクの中では，自力でリカバリーできないほど探索してほしくないという考え方があります．

DeepMindのAIの安全性に関するチームがこうした問題に対してこれまで研究してきた最新の成果について紹介します．一つは深層強化学習のアウトプットに対して，人間が「こちらのほうが良い」といった形でAIにフィードバックすることで，人の嗜好を埋め込めないかの研究をしています．また誤った報酬の仕様によって改悪されてしまった状態になったAIには，どのような情報が役に立つのかを研究しています．さらにどのような形態のAIが中断可能な状態になれるかを研究しようとしています．

現在，大学から産業界，シンクタンクまでAIの長期的な安全性についての研究チームが非常に多く，また国際的になってきています．もし皆さんがこのようなオープンプログラムに参加してくださるのであれば，非常にうれしく思います．

3. AIが人間の仕事を奪うとき

中川：人間は産業革命が起こるまで，多くの労働をしてきましたが，産業革命の後に工場ができ，肉体労働はほぼなくなり，新しい仕事は知的な作業となりました．

強力なAIができると，人間の仕事は奪われるのではないか，そこで人間は何をするべきか，どこで仕事を見つけるのか，どのような仕事を見つけるのか，どのように自分を教育するのか，ということが問題となってきます．

AIの影響を受けない新しい仕事はあるのか，例えばデータなしで判断を要する仕事など，人間のほうがAIよりもクオリティが高い仕事が考えられます．もしくはセールスマンのような対人関係のコミュニケーションを要する仕事やビジネスと新技術のつながりを探す仕事，ニーズが少なく，そのニーズのためにAIを開発することに経済性がないようなサービスをする仕事，またAIを開発する仕事やAIが生み出した成果を人間に説明する仕事が考えられます．しかし，こうした仕事は将来的にAIのほうが効率的に行えるため，人間が行える仕事はなくなってしまうでしょう．

多くの人がこのような観点からベーシックインカムが必要であると主張しています．しかしそれではやりがいをなくし，人間は幸福になれなくなるともいわれます．十分に時間があれば趣味をすればよいかもしれませんが，趣味にはかなりお金が必要で，もはやそのお金を稼ぐための仕事はありません．

また人の仕事がすべてAIに代替され，もはや仕事のプロセスに人間が介入しなくなった場合，例えばもしAIが突然どこかに行ってしまったり，故障してしまったりしたら，人間はどうすればよいのでしょうか．

一方で明るい話もあります．多くの先進国，とりわけ日本では労働力不足の問題があります．この問題に対して，AIとロボットが効果的な解決策になるのではないでしょうか．例えば，自動運転車によって運送業のドライバ不足は解決されるかもしれません．

またAIの下す決定に対して，反論できる権利をもてるように社会を設計する必要があります．そのような政策を先進AI社会に織り込む必要があります．ある意味，我々はAIと戦う必要があります．とはいえ，人間の力は弱いので，AIとの戦いは戦う前に終わらせるか，AIを使ってAIと戦うしかありません．今後AIで発生した問題はAIで解決しなければならなくなるでしょう．

4. 人間とAIのより良いパートナーシップを築くには

ロッシ：AIを我々の社会に取り入れていくことは避けられないものです．そのうえでAIと人間の間に良いパートナーシップと信頼を築くためにはどのようなメカニズムがあるのか，正しいレベルの信頼を築くためにはどうしたらよいかという観点からお話しします．

AIは私達がより良い意思決定をするのを助けてくれます．また，人間がやるべきではないこと，やりたくないことを避けるのを手伝ってくれるものです．そのためAIは自律的に物事をやってくれるものとして捉えるだけでなく，人間と協力してくれるものだという捉え方をするべきだと思います．フォーカスするべき方向性は人間とマシンの共生です．

人間とAIは補完性があります．人間は，正しい質問をすることができますが，その答えを出すには機械の能力で補完してもらうことができます．また，人間は常識に基づいた理由付けを得意としている一方で，統計的推論や統計的処理は機械が得意としているので補完性があります．それぞれ得意としているところが違っているので，機械と人間が一緒になることでより良い意思決定を行うことができます．例えば医療においても医者が最適なAIを使えば，誤った診断をする確率を大幅に下げることができます．

ただし，これには多くの倫理的問題があります．まず，クラコフナ氏が取り上げたように，価値観が一致しているのかという問題は最も重大な問題の一つです．例えば医療におけるAIによる意思決定支援システムは，医師が守るべき倫理的な原則を守る意思決定システムであるべきで，そのためにはAIが倫理原則を理解し，それに従った仕方で医師に発言しなければなりません．一方で，医師が倫理的な原則から外れそうな場合は，警告を出すような形で意思決定を助けることができるようになるかもしれません．

二つ目は信頼に関する問題です．システムが勧める意思決定は本当にベストな意思決定だとどうすれば信頼できるのでしょうか．この問題はAIの説明能力と関係しています．特に医療や司法の分野においては，説明能力というのが必要になってきます．つまり説明をしてくれて，本当にその提案に従ってよいのか信頼させてくれる能力が必要となります．

このように信頼性，人間と価値を一致させる方法，説明能力，それから説明能力に関係してくる透明性，これらをどうやってAIシステムに埋め込むことができるのかということを多くの研究者が現在考えています．

それから短期的な懸念の中に，データバイアスの問題があります．データ量やAIのアルゴリズムだけでなく，元のデータに多様性があるのか，十分に代表性があるかどうか，正しく分布しているのかということも重要になります．

また，どうやって人間の価値観をAIに埋め込んでいくのか，というのも問題です．人間の価値観は文化に特異的であるという話もありましたが，社会的な規範，倫理的な原則，行動原則，こういったものが文化によって異なる可能性があります．対立する価値観がある場合，その対立を解決する必要もあります．

私達は今，AIでもこういった問題を扱っており，具体的なガイドラインをつくろうとしています．例え

ば，透明性とはアルゴリズムとデータを見ることができたらよいというわけではなく，AIが自らやったことを人間に理解できるように説明するということです．また価値観を埋め込んでいくには，データ中心のマシンラーニングアプローチとルールベースのアプローチの二つのアプローチを組み合わせるということが重要であるなどと議論をしています．

機械が現実社会で使われるようになれば，倫理的に行動することが求められるでしょうが，社会に対してどのような影響があるのか，そして人と人のコミュニケーションにどう影響するのか，それから教育はどうなるのでしょうか．また，透明性をもって信頼できる形でAIシステムを開発する必要があります．例えばIBMは企業の責任として，顧客のデータは顧客のものとして扱い続けるということを宣言しています．それからコミュニケーションとコラボレーションを通じて，ガイドラインをつくろうとしています．倫理的なAIをつくろうとすることが必要だと思っています．

5. 総 合 討 議

オヘガティ：根本的にAIの安全性にはどのような問題があると思いますか．また，今後AIの安全性を確保するために何が重要だと考えますか．

クラコフナ：今，長期的なAIの安全性に関して検討している多くの分野，例えば介入できるかどうか，悪い副作用を回避することはできるかなどの研究がかなり根本的なものだと思っています．現在のシステムが将来のシステムにどのように引き継がれるかは研究の進歩によって大きく変わりますが，強化学習など現在の技術は将来の先端的なシステムでも関連性をもち続けると思うので，このようなシステムに関する根本的な研究は将来的にも意味があると思います．

ロッシ：バイアスの問題も重要だと思います．意図せざるバイアスがあるような場合，開発者に対して警告を出すシステムを開発しようとしています．また，既存のAIにどのくらいバイアスがかかっているのか評価尺度に基づいて評価することも考えています．バイアスが本当にあるのか，あったとしたらそれは意思決定に深刻な影響を与えるバイアスなのか，といったことも評価できると思います．

また，私は報酬ベースの機械学習と倫理的なポリシーを組み合わせることにも興味があります．システムが報酬を最大化することだけに左右されるのではなく，付随する倫理的な問題についても学習をするようにしていきたいと考えています．

オヘガティ：中川氏はこれらの問題に対してAIの研究者はどの程度問題を解決することができると考えていますか．また，政治や教育，社会に対してAIの研究者はどういった責任をもつことができるとお考えですか．

中川：バイアスのないデータがAI研究者にとって重要です．それに加えて，公平さが非常に重要となります．もしAIが公平であれば，信頼することができるからです．

オヘガティ：これまで触れてきたのは長期的な問題が主でしたが，例えばバイアスや透明性，信頼の問題は現在にも当てはまる問題でもあります．長期的な問題について考えていると，短期的な問題から目をそらされてしまうかもしれませんが，そういうことは心配していますか．それともこういった問題は一緒に議論すべきでしょうか．私は長期の問題，短期の問題，どちらも重要だと考えています．そして解決策としても短期と長期をまとめて扱ったほうがより効果的だと思っています．

クラコフナ：確かに長期の問題を考えると現実問題を忘れてしまうかもしれません．長期的なことに不正確な警鐘を投げかけるだけの報道もあり，本当に議論しなければいけない問題を取り上げていないと感じます．

私は短期の問題も長期の問題もどちらも非常に重要だと思います．グッドニュースは短期的な問題も長期的問題も研究が増えているということです．

ロッシ：長期的な問題を解決するには短期的な課題にも取り組まなければなりません．メディアの問題は，現実の問題から目をそらしているだけでなく，AIの能力を誇張している点にもあります．

Partnership on AIで学生だけでなく，政府，一般市民，それからメディアへの教育的取組みを考えています．今どういった課題があるのか，そしてどのような限界があるのかについて誰もが知らなければいけないと思います．それは科学的にも，またAIの展開という意味でもです．それによって私達のような科学者が適切な問題に焦点を当てることができるようになります．

中川：汎用人工知能（AGI）ができたとして，私達の期待どおりのことを実現できるのかどうか，真剣に考えなければいけないと思います．フラッシュクラッシュのように分散型AGIが他の情報チャネル，例えば為替レートと情報交換をした場合，予想外のことが起こるかもしれません．このようなことを早く検知するのもAIの仕事になります．AIの不正を検知できるAIも必要となるでしょう．

文責者あとがき

本セッションではいかに人間にとってより良い形でAIを利用するか，いかにAIと共生するか，ということがメイントピックの一つであった．そのためには人間の価値観をAIに埋め込むということ，またAIの透明性・説明能力が求められる．

人間の価値観は社会や文化によって違うだけでなく，個人個人によってもかなり違う点があり，完全にすべて

の人の価値観を一致させてAIに埋め込むことは難しいだろう．将来，できるだけ多くの人の価値観に合ったAIが利用できるように，今現在からさまざまなセクタを巻き込んだ議論が必要だと考える．その点，長期的な問題と短期的な問題を分けて考えるのでなく，同時に考える姿勢が大切であろう．

2018年2月2日　受理

著　者　紹　介

松井　拓海
東京大学教養学部2年.