特　集「人工知能学会・情報処理学会共同企画―第２部「人工知能における人道とは」―」

AIが社会に受け入れられるための「人道」とは？

What Is Socially Permissible“Humanity”for AI ?

栗原　聡
Satoshi Kurihara
電気通信大学大学院情報理工学研究科，人工知能先端研究センター
Department of Informatics and Engineering, The University of Electro-Communications. /Artificial Intelligence eXploration Research Center.
skurihara@uec.ac.jp, http://www.ni.is.uec.ac.jp/

「AI開発における倫理」という話題は，一般社会においてはAIの具体的研究成果よりもはるかに関心が高い．先日，NHKラジオ「先読み！　夕方ニュース」での特集「AI社会の未来とリスクは？」に出演させていただいた際も，通常の3倍もの意見・コメントが寄せられるなど，その関心の高さを実感したしだいである．今回の本学会と情報処理学会との連携企画において，第2部では一人の著者が両方の学会誌に寄稿し，情報処理学会誌には，各自の専門領域の動向について，そして本学会誌には「人道」というお題にて自由に意見を書く，ということとなった．第2部担当の著者の方々が，それぞれ抱く「人工知能における人道」はどのようなものなのか？　このような企画が採用されるのも本学会ならではの味であろう．AIが人道的に動作するのであれば人はAIを信頼し，安心して付き合うことが可能になるのであろうか？　そもそも人道的AIは実現可能なのであろうか？

Microsoftのチャットボットである Tayが不謹慎な発言をするようになってしまい，すぐに運用が停止されてしまった話題や，ハンソンロボティクス（Hanson Robotics）のAIロボット「ソフィア」が同じく不謹慎な発言をした例など，AIが社会に対して不安を与える事象が起き始めている．Tayとソフィアは人道性の欠如したAIなのであろうか？　両者ともシステム自体に不謹慎な発言をする仕掛けが組み込まれていたわけではなく，不謹慎な発言をするように人が学習させたに過ぎない．人道的でなかったのは学習させた人間のほうであろう．しかし，人道的とはいえない行動が可能なシステムであることは間違いなく，そのような行動が発動することがない仕掛けをあらかじめシステムに組み込むといった防御策が必要であることが今回の出来事でより明確化された．

著者はこの7月にニューヨークにて開催された汎用AIの国際会議AGI 2016[*1]に参加し，ソフィアの実物とその対話デモを見ることができた．

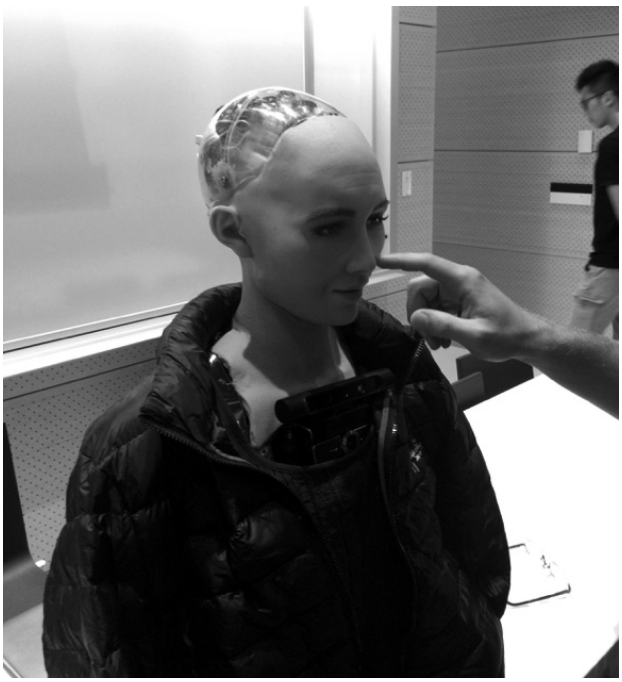

図3　Hansonロボティクスのソフィア

対話は対話はけっしてかみ合っていたとはいえず，にもかかわらず流暢にしゃべるので，ちぐはぐ感がさらに強調されてしまっていた．場の空気を読まずにひたすらしゃべる人のような印象である．この状況であれば例の不謹慎な発言も，会話のコンテキストを理解しての発言である可能性は極めて低く，単に不謹慎な発言をしたとしても，それはそのような発言がシステムにあらかじめ記憶されており，偶然発話されたに過ぎず，人道について議論する以前の問題であると感じた．

Siriにせよ，チャットボットにせよ，現在利用可能な対話システムと人同士のような生きた会話ができると実感できる人はいないのではないだろうか．音声対話システムをはじめ，さまざまな用途で音声を用いたナビゲーションシステムが導入されてはいる．ペッパーや，最近発表されたロボホンなども音声による対話ができることが特徴である．しかし，現状は，人同士のような違和感

*1 http://agi-conf.org/2016/

のない自然な対話ができるわけではなく，文字でのやり取りを音声に置き換えたレベルであったり，質問に対する回答という，定型的なタスクがほとんどであろう．対話システムが利用するデータ量は，すでに我々人間の知識量を大きく上回っているであろうし，それにもかかわらず，対話システムとのやり取りにおいて，人同士のような生きた対話ができない大きな理由がある．それは，現在の対話システムは目的指向性を有してはいないことである．スマホの音声対話システムに対して「喉が渇いた」と話しかければ，必ず直近のコンビニや自販機の場所が回答として返ってくる．しかし，人同士の場合「今は我慢して！」などと返答する場合がある．直近の自販機には水以外の高カロリーなジュースしかなく，相手の糖分取り過ぎによる健康への悪影響を防ぐための発言である．つまり，この場合「相手の健康を気遣った」，別の解釈をすれば「相手の幸福度を向上させたい」という目的を達成するために「今は我慢して」という発言をしたのである．このような何気ないやり取りが，お互いの信頼感を生み出している．相手への気遣い以外にも，「その場の雰囲気を維持したい」という目的や，自らの欲望を達成するための発言など，我々はさまざまな目的をその場その場の状況で適切に選択し相手との会話や振舞いを行っている．現在の対話システムには，このような目的指向性がなく，単に与えられた質問に解答するのみであることから，そもそも人同士のような会話の成立は不可能なのである．そして，対話に限らず，人に寄り添うAIを実現するためには，AIが人に能動的にインタラクションを実行するための目的指向行動選択能力が必要となる．

加えて，さらに足りないのが，対話システム自体が対話相手のモデルを想定し，想定したモデル駆動にて発言を生成する機能をもたないことである．ところで，我々の意識や自我は明らかに脳神経細胞ネットワークが創発させるダイナミクスであるが，その意識は自分を創発させている神経細胞の挙動を知覚することができない．およそ2 000億個にもなる脳神経細胞が総延長100万kmにもなる軸索や樹状突起から構成させる超大規模複雑ネットワークの振舞いとして，発話が生成される．発話された言葉は，それを聞く相手の脳の超大規模複雑ネットワークに反応を起こすためのトリガーに過ぎない．つまり，発言される言葉のみに着目しても，「空気を読んだ会話」，「阿吽の呼吸的会話」ができるわけがないのである．そして，仮に対話相手のモデルを構築する機能を実装したとしても，次に問題となるのは，マルチモーダル性だと考えている．深層学習法の欠点として，学習に大量のデータを必要とする点がよく指摘される．では，我々はどうであろうか？　幼少期にネコを見分けられるようになるまでに100万匹ものネコを見てはいまい．恐らくは数匹であろう．だからといって人は少ないデータで学習できると言い切れるであろうか？　人は1匹のネコであっても三次元動画としてその動き方を捉え，同時に鳴き声や触った場合にはその柔らかさ，そしてネコを見たときの情景など，五感を通して入るすべての情報を関連させてネコの概念を獲得している．よって「ネコ」と言われたときに各自が想起するネコはネコの真正面の顔ではなく，それぞれが経験した情景として想起される．しかし，深層学習においては，ネコの顔の二次元画像しか与えられていない．よって，大量のデータが必要になると思えば納得もできよう．つまり，対話システムにおいても，対話相手の人のモデルを構築する際，人の発話情報のみではモデル構築は極めて困難であり，表情やその場所の情景など，さまざまな情報をお互いにネットワーク化しつつ人のモデルを学習する必要があろう．無論，初めて対面した人とのやり取りがぎくしゃくしたものになるのは，人同士であっても同様である．しかし，学習が進むにつれ，その場に合致したコンテキストと，目的指向性により，生きた会話が可能になるのではないだろうか．

まとめると，現在の対話システムやインタラクションロボットにおいては目的指向性が存在せず，能動的な動作能力が組み込まれていない以上，人道的という観点からはそれ以前の問題であり，人道的でない発言なども，それを学習させた人間側の問題であるものの，そのような発言をしないような機構を入れ込む必要があろう．そして，目的指向性などが組み込まれ，能動的に対話可能となる進化したAIにおいては，目的指向性の部分に人道的に行動する規則をいかにして入れ込むかが鍵となりそうである．まさにSF映画のような話であるが，指数関数的に進化しつつあるAIをはじめとする科学技術においては一見，荒唐無稽と思われるくらいの予測のほうが現実になる状況になってきたと最近実感する．まさに人道的AIについて真剣に考えるときが来たのだと思う．

2016年7月24日　受理

著　者　紹　介

栗原　聡（正会員）

慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．NTT基礎研究所，大阪大学大学院情報科学研究科・産業科学研究所を経て，2012年より電気通信大学大学院情報システム学研究科，2013年より情報理工学研究科教授．同大学人工知能先端研究センターセンター長．大阪大学産業科学研究所招聘教授．ドワンゴ人工知能研究所客員研究員．内閣府科学技術・学術政策研究所客員研究官．博士（工学）．人工知能，複雑ネットワーク科学などの研究に従事．著書『社会基盤としての情報通信』（共立出版，2000）．翻訳『群知能とデータマイニング』，『スモールワールド』（東京電機大学出版局，2012，2006）など．本学会理事・編集委員長などを歴任．