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」―」

人道知能と汎用人工知能

Humane Artificial Intelligence and Artificial General Intelligence

市瀬 龍太郎 国立情報学研究所
Ryutaro Ichise National Institute of Informatics.
ichise@nii.ac.jp

1. 人 道 知 能

「人道」という言葉は，あまりにも大きく漠然としている．本稿において，人道知能がどのようなものを意味するか，まず初めに議論する必要があるだろう．「人道」という言葉を「広辞苑」で引くと「人のふみ行うべき道．人の人たる道．人倫」と記述されている．大半の人工知能学者がつくろうとしている人工知能も，人のふみ行うべき道に沿った人工知能ではあろう．しかし，人のふみ行うべき道とは，何であろうか．

ここで，「人道」の意味に出てくる「人倫」という言葉に着目し，再度「広辞苑」を引くと，意味の中に「人と人との秩序関係」という記述がある．人道知能とは，「人と人との秩序関係」を守る人工知能と考えると少しわかりやすくなったように思われる．そこで，本稿では，「人道知能」とは，「人と人との秩序関係」を守る人工知能と考え，議論を進めていくことにする．ただし，ここでの「人道知能」は，警察のような秩序関係を守らせる主体ではなく，あくまでも，人と人の中にいる秩序関係を守る主体である．

2. 人と人との秩序関係を守る人工知能

人道知能を設計するためには，人と人との秩序関係がどういうものであるのかを何らかの形で人工知能のほうに実装する必要がある．実装の仕方としては，大きく分けて二つの考え方があるであろう．一つ目が知識として人工知能プログラムに組み込む方法，もう一つが学習により，人工知能プログラムに秩序関係を取り込む方法である．本章では，それぞれについて考察する．

2·1 人道的知識の定式化

人と人との秩序関係を知識として人工知能に取り込むためには，何らかの形で「人と人との秩序」を記述することが必要となる．ここで，「秩序」という言葉を再び，「広辞苑」で引いてみると，「社会などの規則立った関係」と記述されている．規則ということであれば，一見，知識工学的に知識を記述し，人工知能システムに取り入れることが可能なように考えられる．しかし，人道的な判断においては，人間においても判断に迷うようなものが多くあり，人道的に何が正しいものであるかを規則として明示化することは大きな困難を伴う．そのような代表的な問題として，トロッコ問題というものが知られている．走っているトロッコにトラブルがあり，止められない状況を想定する．そこで，ポイントを操作することで犠牲者を変えることができる．しかし，誰かが犠牲者となる状況は変えられない．その場合に，どのような対応をするべきであるか，というのが，問題の大枠である．人道的には，人に危害が及ばないようにするべきであろうが，避けられない状況の場合に，ポイントを操作せずに二人を犠牲にするほうを選ぶのと，ポイントを操作し一人を犠牲にするほうを選ぶのと，どちらが人道的であろうか．もし，犠牲者の少ない一人のほうを選ぶとして，その一人が妊婦でお腹の中に子供がいたら，それは人道的なのだろうか，など，規則化するには非常に大きな困難が伴う．その一方で，人工知能を使った自動運転が現実的になってくると，この状況は，けっして絵空事の話ではなく，自動運転車に搭載する人道知能を考えることが近未来のこととして必要になる．Bonnefon らは，自動運転車で被害者が出ることが避けられない状況の例と，その状況における人の判断がどうなるかを調査することによって，人道的な判断について議論を行っている [Bonnefon 16]．その議論の中では，自動運転車のために，どういった状況でどういう人道的な判断を行うかの規則化は必要であるが，規則化により自動運転車自体が敬遠される逆効果も及ぼすとされている．そのような作用まで考慮すると，人道的な知識を定式化し，明示的に人工知能に入れることは，非常に大きな課題となるであろう．

2·2 人道的判断の学習

機械学習は，さまざまな事例を用いることで人工知能の判断を改善することができる．そのため，人道知能をつくるうえでも大きな威力を発揮することが期待される．人と人の間に置くことにより，どのような秩序関係が成り立っているのかを学習することができるからである．しかし，学習機構というのは，周りの環境をモデル化するため，環境からの影響を大きく受ける．そのため，人と人との秩序関係をどこまで汎化するのかが大きな課題となる．例えば，マイクロソフトの作成した対話エージェント Tay が不適切な発言をし，謝罪をすることになった事例 [Lee 16] は，人道知能を作成するうえでの大きな教訓を残していると考えられる．Tay には，学習機

構が備わっており，環境（対話）から学習ができるようになっていた．その対話を通して，対話を行っていた特定の人々の秩序（差別的な発言を行う）を学習しており，人と人との秩序関係を学習できていたとも考えられる．ただ，学習をした環境が狭く適切でなかったために，人道から外れた人工知能になってしまったといえるであろう．つまり，学習により，人道知能を作成するには，人類全体を対象として汎化を行わなければならないという課題がでてくるといえる．

3. 人道的な汎用人工知能

汎用人工知能とは，異なる領域において多様で複雑な問題を自律して解いていく人工知能である [市瀬 16]．人道的な汎用人工知能を構築する際の問題として，汎用人工知能に特徴的となるメタ学習に関して，本章で考察する．

メタ学習とは，人工知能システムが自己の変更を行い，性能を改善する能力である．自己の変更とは，学習パラメータの変更のみならず，学習手法自体の変更や，内部の機構自体の変更を含み，一般的な機械学習が取り扱う学習のメタなレベルでの変更となる．汎用人工知能においては，多様な領域に対応するために，自己を適応化する必要があり，そのためメタ学習が必要となる．自己を改変することにより，人工知能がより賢い人工知能になることができる．そのような汎用人工知能を想定したときに，人道知能を実現するには，どのようなアプローチがあるだろうか．

一つの考え方として，環境から与える情報を制御することで，人道知能を実現するという考え方がある．これは，強化学習などの枠組みを主に想定しているもので，人間が望ましい報酬や学習データを与えることで，人道知能を実現するアプローチである．しかし，このアプローチは，前章の学習で議論したアプローチ以上に困難を伴う．なぜならば，汎用人工知能においては，内部を書き換えて自己を改変することが可能となるため，報酬や学習データ自身を書き換えてしまうことができるからである．そのようなことを防ぐ方法を考えていくことが必要となる．

別の考え方として，解釈型人工知能（Constructionist AI, 構成主義人工知能（Constructivist AI）とは別の概念）という考え方 [Thorisson 16] がある．解釈型人工知能では，学習することを限定し，人手で人工知能を設計することにより，人道から外れた人工知能にならないようにする考え方である．しかし，そのことにより，汎用人工知能に必要となる自律性が限定されてしまうという課題も生ずる．この解釈型人工知能の考え方は，人道知能を構成するための一つの考え方であり，アイスランド知的機械研究所（IIIM）では，殺人ロボットなどを防ぐために，平和的研究開発のための倫理方針 [IIIM] でこの考え方を取り入れている．

4. 社会と人道知能

ここまで，人道知能とは何か，いかにして実現するかについて，議論をしてきた．しかし，議論を通して，人道知能を実現するのは，従来の工学技術とは違う側面があることが浮彫りになってくる．人道知能とは，「人工知能をつくる人の人道」,「人工知能そのものの人道」,「人工知能を利用する人の人道」の三つが複合されて実現できるものであるといえる．まず，人工知能をつくる人自身が人道を理解しながら，人工知能をどう設計するかを考えていく必要がある．そして，設計された人工物としての人工知能そのものが，人道に配慮しながら動作するようにする必要がある．さらに，人工知能に学習のための知識を与えたり，利用法を決めたりする利用者が人道を考えながら利用する必要がある．この三つがそろい，初めて人道知能がつくれるようになる．

人道知能というのは，単なる工学技術だけで構成できるものではなく，社会的な合意があって初めて構成できるものである．そのために，社会を巻き込みながらつくっていく必要があり，人工知能研究の新たな挑戦となるであろう．

◇ 参 考 文 献 ◇

[Bonnefon 16] Bonnefon, J.-F. Shariff, A. and Rahwan, I.: The social dilemma of autonomous vehicles, *Science,* Vol. 352, No. 6293, pp. 1573-1576（2016）

[市瀬 16] 市瀬龍太郎：汎用人工知能の現状と展望，情報処理，Vol. 57, No. 10, pp. 960-961（2016）

[IIIM] Icelandic Institute for Intelligent Machines, Ethics Policy, http://www.iiim.is/ethics-policy/

[Lee 16] Lee, P.: Learning from Tay's introduction, Official Microsoft Blog（2016）, http://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

[Thorisson 16] Thorissson, K. R.: Kristinn R. Thorisson RU Home（2016）, http://www.ru.is/faculty/thorisson/

2016 年 8 月 9 日　受理

著 者 紹 介

市瀬　龍太郎（正会員）

国立情報学研究所情報学プリンシプル研究系准教授．総合研究大学院大学准教授．博士（工学）．知識処理，機械学習の研究に従事．本学会編集委員会副委員長，汎用人工知能研究会主幹事．電子情報通信学会，情報処理学会各シニア会員．AAAI，日本認知科学会各会員．