SF の分析を用いた未来のエージェント像の検討

Investigation of Future Agents using Analysis of Science Fiction

大澤博隆 [1,2] 長谷敏司 [2] 宮本道人 [1] 西條玲奈 [3] 福地健太郎 [4] 三宅陽一郎 [5]

Hirotaka Osawa[1,2] Satoshi Hase[2] Dohjin Miyamoto[1] Reina Saijo[3] Kentaro Fukuchi[4] Yoichiro Miyake[5]

[1] 筑波大学
[1]University of Tsukuba
[2] 日本 SF 作家クラブ
[2] Science Fiction and Fantasy Writers of Japan
[3] 京都大学
[3] Kyoto University
[4] 明治大学
[4] Meiji University
[5] 日本デジタルゲーム学会
[5] Digital Games Research Association Japan

Abstract: 我々は、人間と共存する将来の AI エージェントシステムの新しいビジョンを創造するため、SF における AI エージェントの利用法を、プロの SF 批評家および作家と共同で調査した。近年、発達する人工知能が仕事における人間の役割を奪う危険性が議論されており、その際に一部の SF が引用されることがある。しかし実際の SF においては、人間と衝突するだけでなく互いに補完する多くのエージェントシステムが提案されている。本研究では、これらの多様な SF の傾向を統計的に分析し、エージェント設計におけるアイデア源として用いることを提案する。本研究では、AI エージェントの多様性、AI エージェントの社会的側面、AI エージェントによる人間知能の拡張という 3 つの方針に基づき、AI と人間の関係を特徴的に記述する 115 の SF 小説を、専門家の手を借りて選んだ。次に階層クラスタ分析と主成分分析を用いて、SF における AI の特性を表す 11 因子を分析した。その結果、AI エージェントには人間型、機械型、補助型、設備型の 4 つのタイプがあり、知能と人間らしさの 2 つの主要な次元があることが示唆された。筆者らはこの分析に基づいて、人-エージェント相互作用に関する将来の設計のために注意すべきステレオタイプと、将来設計に参照可能なビジョンを検討する。

1 はじめに

HAI 分野における SF の影響は顕著である。SF は、HCI からロボティクス[1]–[5]までのいくつかの研究分野でイノベーションを生み出す手法として注目されている。Shedroff らは、SF がデザイナーや研究者に影響を与える要素として (1) インスピレーション、 (2) 期待、 (3) 社会的文脈、 (4) 新しいパラダイムの 4 つを定義した[6]。

SF は、特に人間と AI エージェント間の相互作用を設計する、人間-エージェント相互作用(HAI)の分野において重要な役割を担っている。例えば、Isaac Asimov のロボティクス SF は、人間とエージェントを設計するための規範の一つである。彼のロボット工学の三法則[7][8]は、フィクションを超越し、現実世界の人間-エージェント設計に影響を与えてきた。一方で、人工知能 (AI)エージェントが仕事を奪ったり、制御不能な人工知能 (AI) によって破滅を招くなど、SF が描いてきた暗い未来像が懸念されており、例えば映画『ターミネーター』に登場するスカイネットは、人工知能の負の未来像の 1 つと言われることが多い[3]。また、SF のステレオタイプが我々の視点をより保守的にしているという SF の影響に対する批判もある。Robertson は、日本政府の将来像が古

典 SF の影響を無批判に受けており、性差別的なイメージを持っていることに問題を提起している[9]。

SF ステレオタイプの影響を回避しながら、SF を HAI 関係のヒントを得るための題材とするため、本研究では日本 SF 作家クラブの助けを借りて、SF でエージェントがどのように扱われるかを調査した。

2 エージェントの選択基準

まず作品に登場するエージェントの選び方が恣意的にならないように、選択基準を作成した。SF でのロボットの使用に関する以前の Mubin らの研究は、SF レビューの統一基準に基づいて作品を選択することが重要であることを示唆している[10]。本研究では、SF をレビューする際の基準として、特定の組織による殿堂入りを用いているが、SF に登場する AI エージェントはロボットよりも多様であり、既存の基準だけでは追いつかない。また、初期の SF の AI エージェントは、AI が定義される前はそもそも AI と呼ばれていなかったため、単語だけで収集することはできない。

基準の作成にあたっては、SF 作家やファンタジー作家のエキスパート 15 名とオンラインでディスカッションを行い、海外 SF 作品、国内 SF 作品、コミック、若手小説、映像作品、ドラマのそれぞれの専門分野を持つエキスパート 7 名(評論家 6 名、作家 1 名)を選定した。彼らと著者(2 人の科学技術者と哲学者)との半日の対談をもとに、SF に登場する AI エージェントの選定基準を以下のように設定した。

SF に記述されている AI エージェントについて、以下の 3 つの異なる役割があることを念頭に置いて選択基準が作られた。

1. 異星人の知性の可能性を追求した作品。プログラム、ロボット、異性の知性など、様々な形の知性を描き出している。Greg Egan の「スティーヴ・フィーヴァー」に登場する群知能ナノマシンの Stevelets など。
2. 知能の社会的側面を追求した作品:エージェントの知能の詳細な実装が物語に記述されていなくても、AI エージェントとの社会的相互作用が物語の要因となる作品。星新一の「ボッコちゃん」など。
3. 人間の知能を人工的に拡張する可能性を追求した作品。テーマは、人間とロボット/マシンの高度なインターフェース、人間の拡張、インターネット、ソーシャルネットワークを通じた人間の認知能力の拡張。操作者の能力を伸ばす戦闘機エージェントの神林長平の「雪風」など。

上記の基準で選択した SF 上でエージェントを分類するため、以下の 20 項目を収集した。

1. 定量的項目 11 件 : 製作者の規模、独立性、友好性、汎用性、自意識、群集性、ネットワーク接続性、言語能力、学習能力、物理性、人間型類似性
2. 定性的項目 9 件 : エージェント名、作品名、出版年、メディア (小説、漫画、映画、演劇)、発行国、タスク、エージェントのコミュニケーション手法、素材、エネルギー源

SF の専門家に調査を依頼するにあたり、最終的に我々は以下の選択基準を選んだ。

- 多様性:出版された時代や媒体が特定の分野に偏っていてはならない。
- インパクト:社会的に大きなインパクトのある作品を集める。社会的影響は小さいが特徴のあるものも収集するのが適切である。
- 独自性:類似した特性を持つ AI エージェントの場合は、オリジナル作品を挿入する。1 つの作品に複数の AI エージェントが登場する場合、最も特徴のあるものを優先的に集める。

3 データ収集と分析

専門家による相互品質チェックの後、115 の AI エージェントを収集した。作品の平均年は 1981 年(S.D.26.8 年)。最古の作品は、1912 年にヤロスラフ・ハシェクが書いたサイボーグ作品、最新の作品は 2019 年に柴田勝家が書いた粘菌制御によるバイオ AI エージェント「ヒト夜の長い夢」の少女 M で、出版国は日本 55 カ国、アメリカ 52 カ国、イギリス 3 カ国、ポーランド 2 カ国、チェコ 1 カ国となる。また、全世界で同時に出版されている作品も 2 つあった。最初の媒体は、小説が 93 本、コミックが 12 本、映画(そのうちの 2 つはアニメーション)が 7 本、演劇が 3 本であった。1945 年以前は 14 件、第二次世界大戦後はインターネットの普及以前の 1995 年までは 64 件、1995 年以降は 37 件と、年代ごとに多様な分布を示している。

定量的項目 11 因子について階層的クラスタ分析を行った。各要素間の距離はユークリッド距離で測定し、Ward 法で分類した。クラスタツリーを、特徴点における距離 10 に基づいて 4 つのクラスタに分割した。また、11 因子について主成分分析を行った。その結果、第二主成分まで 41.4%、第四主成分まで 64.4%の情報量が抽出された。高い寄与率を持つ最初の軸(23.0%)は知能とラベルされ、二番目の軸(18.4%)は人間性とラベルされた。本論文では、分析のため

に各エージェントの知能と人間性を主に使用する。

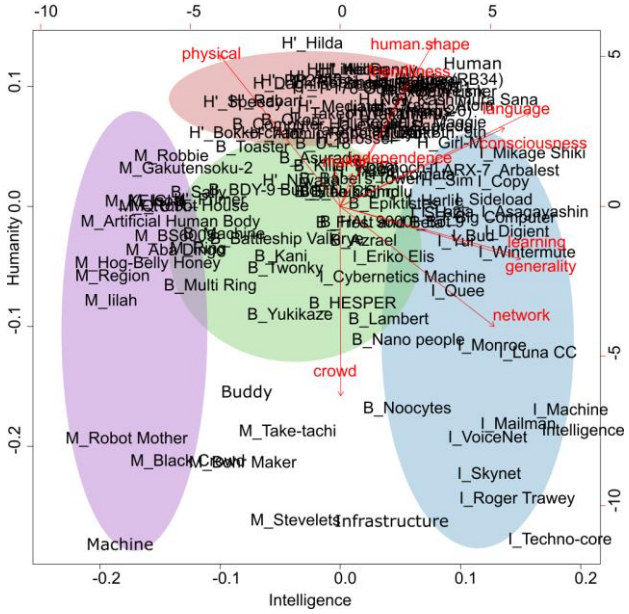

図 1: 主成分分析による知能と人間性に関する図及び SF に登場するエージェントの 4 つのクラスタ

4 つのクラスタは知能レベルと人間的レベルに基づいて 4 つのタイプに大別された。4 つのクラスタの特性に基づいて、「人間型」、「バディ型」、「機械型」、「インフラ型」というラベルを付けた。いくつかの要因(矢印)は相互に関係していることがわかった。例えば、人間的な形と親しみやすさ、そして意識と言語能力の両方が知性と人間らしさの向上に貢献する。一般性、学習、およびネットワーク接続性の向上は、知性の向上に寄与するが、人間性の低下にもつながる。群集の要素は知性には寄与しないが、人間性を減少させる。身体的要因は人間性を高めるが、知性を低下させる。独立性とメーカーはどちらの軸にも寄与しない。電力が最大のエネルギー源であり、その他のエネルギー源は多様であり、大きな傾向は見られなかった。

3. 結果

人間型エージェントは、4 種類の型の中で最も多い。この型は、中程度の一般性(.47(SD .33))、高い意識(.84(SD .30))、高い言語技能(.95(SD .15))、中程度の学習技能(.53(SD .45))、高い物理的外見(.97(SD .13))、と高い人間的外見(.93(SD .16))がある。手塚治虫「鉄腕アトム」のアトムや、アイザック・アシモフの「われはロボット」のいくつかのロボットなどが代表的な例である。AI エージェントによって実行されるタスクでは、家事が最も一般的なタスク(11 件)であり、外部環境における肉体労働(5 例)がそれに続く。人間型エージェントは、人間と同様に、独立して行動し、環境から学び、一般的な仕事を行い、社会の一員として行動する。一般的に、それらのほとんどは人間の比喩として扱われる。この型は図 1 のように比較的狭い領域に集中しているが、これはヒトのイメージが一般的であるためと考えられる。

機械型では、人間より知能の低い AI エージェントが多い。人間型とは異なり、この型は低い一般性(.16(SD .32))、低い意識(.11(SD .26))、低い言語能力(.14(SD .34))、低い学習能力(.18(SD .30))および低い人間類似性(.11(SD .26))という特徴がある。モーリス・ヒューギの「機械ねずみ」のロボット・マザー、安部公房の「第四間氷期」の KEIGI-1 などが代表的な例である。実行されるタスクは、ベビーシッターから武器にまで及ぶが、多くは学習機能を持たない。これらのエージェントは問題に特化した知的な自動機械ではないとして物語に登場し、その融通性のなさがしばしば人間社会に損害を与える。

バディ型エージェントは、人間に依存し、意識的で、協調的な仕事エージェントである。人間型と類似しているが、汎用性の低さ(.29(SD .43))、言語能力のわずかな低さ(.73(SD .32))、人間への非類似性(.00(SD .00))の点が異なる。アーサー・C・クラークの「2001 宇宙の旅」の HAL9000(宇宙船エージェント)、神林長平の「雪風」の雪風(戦闘機)、福田己津央の「新世紀 GPX サイバーフォーミュラ」のアスラーダ(スポーツカー)などが代表的な例である。バディ型エージェントは、ツール型の形状を有し、各ツールに対して特定のタスクを実行し、非言語入力が多い。エージェントによって実行されるタスクの中で、最も一般的なタスクは軍用(8 例)であり、次は自動操作(4 例)であった。

インフラ型のエージェントは物理的にあまり活発でなく、ネットワーク接続性と言語能力を持ち、社会インフラとしてしばしば使用される。人間型とは異なり、この型は高いネットワーク接続性(.96(SD .13))、わずかに低い物理的外観(.62(SD .37))および中程度のヒト型(.47(SD .46))により特徴づけられる。ジェームズ・キャメロンの「ターミネーター」のスカイネット、ウィリアム・ギブソンの「ニューロマンサー」の冬寂(ウィンターミュート)、長谷敏司の「BEATLESS」のレイシアなどがその代表例である。AI エージェントによって実行される最も一般的なタスクは、施設管理(15 件)であった。インフラ型テントのイメージは、第二次世界大戦後、主にコンピュータ技術や通信技術の発展によって生まれたと考えられている。平均発行年は 1996 年(SD 21.8)と他の

年より新しい。

4. 考察

人間型と機械型は、SF に登場するエージェントのステレオタイプであると思われる。人間型エージェントは異文化の人間のモチーフとして使われてきた、SF の機械型エージェントは制御不能な機械のモチーフとして使われてきている。特にカレル・チャペックの R.U.R.のように「ロボット」と名づけられた人間型エージェントは、家事と労働に従事する奴隷と人種差別を反映している。一方で機械型エージェントは、ルールによって制御される融通の効かない愚かさから、制御不能問題を引き起こす。
これらは文学上のテーマとしては興味深いが、技術的に現実的なイメージではないことに留意する必要があると思われる。また、これらのフィクションが HAI にステレオタイプを誘発することも懸念される。HAI 研究の目的の 1 つは人間に似たエージェントの作成であるが、今日の AI エージェントは人間ほど知的ではないし、かといって人間が確認可能なルールに従う機械でもない。商業化されたエージェントの設計やプロモーションには、人間のようなイメージがよく使われるが、イメージの使いすぎには注意が必要ではないだろうか。

将来の HAI 設計のための新しいアイデアを得るには、バディ型およびインフラ型エージェントがより重要であると考える。バディ型エージェントは人間のようなものではなく、人間と協調して作業を行う。これらの SF は、自動運転における人間と AI エージェント間の役割分担を含む、統一された作業システムとして、AI の意思決定と人間の意思決定をどのように妥協するかという意思決定問題を扱う。極端な条件における人ーエージェント間の相互作用問題として扱うことが可能である。またインフラ型は、コンピュータの発展とともに登場した新しいイメージであり、多くの新しいアイデアが存在する分野である。例えば、長谷敏司の「BEATLESS」は技術的特異点後の AI 世界を描いており、Lacia はヒューマノイド・インターフェースとして描かれている。Beatless は今後の HAI 設計で言及される SF の一つと考えられており、AI エージェントを社会に導入する際の倫理的問題、Fogg ら[11]による説得工学の倫理的問題、Affective Computing[12]における社会的要因の性差の問題、及び対応事例を実験的に述べたものと捉えることができる。

5. 結論

HAI 研究に影響を与えてきた SF において、AI エージェントがどのように描かれているかを調査・分析した。その結果、HAI 研究者が SF を参照する際に知っておくべきステレオタイプが明らかになり、今後の HAI の設計に役立つ SF の活動領域が明らかになった。本研究の作品は日米のものが多く、その多くは小説である。日本の映画作品の多くは小説に基づいている。しかし、関連研究には視覚作品の影響に関する多くの研究が含まれており、より多くのメディアを含むように研究を拡大したい。

謝辞

本研究は JST RISTEX「人と情報のエコシステム」内のプロジェクト「想像力のアップデート:人工知能のデザインフィクション」の助成を受けた。また、データ作成に協力いただいた日本 SF 作家クラブに感謝したい。

参考文献

[1] M. Kurosu, “User Interfaces That Appeared in SciFi Movies and Their Reality,” Springer, Cham, 2014, pp. 580–588.

[2] A. Marcus, B. Sterling, M. Swanwick, E. Soloway, and V. Vinge, “Sci-Fi @ CHI-99: Science-Fiction Authors Predict Future User Interfaces,” *Proc. CHI 1999*, no. May, pp. 95–96, 1999.

[3] O. Mubin *et al.*, “Towards an Agenda for Sci-Fi Inspired HCI Research,” in *Proceedings of the 13th Int'l Conf. on Advances in Computer Entertainment Technology - ACE2016*, 2016, pp. 1–6.

[4] P. Nagy, R. Wylie, J. Eschrich, and E. Finn, “Why Frankenstein is a Stigma Among Scientists,” *Sci. Eng. Ethics*, vol. 24, no. 4, pp. 1143–1159, 2018.

[5] J. Tanenbaum, K. Tanenbaum, and R. Wakkary, “Steampunk as Design Fiction,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1583–1592.

[6] N. Shedroff and C. Noessel, “Make it so: learning from sci-fi interfaces,” in *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI ’12*, 2012, p. 7.

[7] I. Asimov, *I , Robot*. 1970.

[8] S. L. Anderson, “Asimov’s ‘Three Laws of Robotics’ and machine metaethics,” *AI Soc.*, vol. 22, no. 4, pp. 477–493, 2008.

[9] J. Robertson, “Gendering Robots:Posthuman Traditionalism in Japan,” in *Recreating Japanese Men*, 2011, pp. 277–303.

[10] O. Mubin, K. Wadibhasme, P. Jordan, and M. Obaid, “Reflecting on the Presence of Science Fiction Robots in Computing Literature,” *ACM Trans. Human-Robot Interact.*, vol. 8, no. 1, pp. 1–25, Mar. 2019.

[11] B. J. Fogg, “Persuasive Technologies,” *Commun. ACM*, vol. 42, no. 5, pp. 26–29, 1999.

[12] R. W. Picard, *Affective Computing*. MIT Press, 1997.