第一工科大学研究報告
第37号（2025）pp.51-55

AI・ロボットに暴言を吐いてもいいか？

寄田明宏　竹下俊一　久保田直行

第一工科大学　工学部　機械システム工学科（〒899-4332 鹿児島県霧島市国分中央 1-10-2）

第一工科大学　共通教育センター（〒899-4332 鹿児島県霧島市国分中央 1-10-2）

東京都立大学　システムデザイン研究科　機械システム工学域（〒191-0065 東京都日野市旭が丘 6-6）

Is it okay to abuse AI/robots?

Akihiro Yorita, Shunichi Takeshita, Naoyuki Kubota
1-10-2, KokubuChuou, Kirishima, Kagoshima
6-6, Asahigaoka, Hino, Tokyo

Abstract: In workplaces that require emotional labor, such as nursing care and customer service, verbal abuse and harassment have become a social problem. AI robots have been developed and will be used in the service industry and nursing care in the future, but there is a possibility that violence and verbal abuse will be used against AI and robots. However, there has been insufficient discussion about whether it is acceptable to verbally abuse AI and robots. In this study, we will consider what impact verbal abuse against AI and robots may have on humans and society, referring to examples from robot ethics and science fiction.

Key words: Ethics, Harassment, Science Fiction

1. はじめに

顧客が従業員らに怒鳴りつけたり，不当な要求をしたりするカスタマーハラスメント（カスハラ）が社会問題となっている． 被害はサービス業や役所に限らず，医療現場でも深刻化しており，患者や患者家族から医療従事者へのカスハラをペイシェントハラスメント（ペイハラ）と呼ぶこともある．「ペイハラ」とは，患者やその家族による医療従事者への暴言や暴力，迷惑行為のことをいう．厚生労働省の主導で平成30年度に実施された「介護現場におけるハラスメントに関する調査研究事業実態調査」によると，今までに利用者から何らかのハラスメントを受けたことがある職員は4～7割だった[1]．また，利用者家族からハラスメントを受けたことがある職員は1～3割と報告されている．サービス形態によってこの割合は異なり，割合が高い順に，介護老人福祉施設が71%，認知症対応型通所介護が64%，定期巡回・随時対応型訪問介護看護が61%となっている．

カスハラやペイハラへの対策は，対応する側のスキルを上げるしかないと言われている[2]．現場でトラブルが起こり「責任者を出せ!」という状況になると，店長は対応しなければならないが，それでも解決に至らないこともある．これを避けるために，組織の規模に関係なく『クレームなどのトラブルに強い組織づくり』が必須となっている．ペイハラに関しては，感情労働に対して感情対処能力を高めることでメンタルヘルスケアも行なっている[3]．

ムーンショット型研究開発制度[4]のプロジェクトでは，人工知能技術（AIT）とロボティクス技術（RT）を融合したAIロボットを社会インフラとして普及させることで，すべての人が社会に積極的に参加できる活力ある社会の実現を提唱している．同プロジェクトでは，スマーターインクルーシブソサエティ（SIS）を提唱し，2050年までにSISを実現するためのAIロボットの開発を進めている．

これまでに，介護支援デジタルプラットフォームを開発してきたが，AITの応用として，よりスマートな共生社会の実現に向けた共進化AIの開発を進めている．共進化AIでは，自己効力感に基

づく支援を司る知能ネットワークを形成し，その知能をクラウドサービスなどを通じて活用していく予定である．共進化AIは，ハードウェアに依存せず，いつでも，どこでも，誰でも利用できるサービスとして，スマートフォンなどの身近な情報端末上で利用できるIaaS（InfraStructure as a Service）などのサービスの形で社会実装することが検討されている．

このようなAI ロボットが人間に代わってサービスを提供する場合に，人間からハラスメントを受けることが想定される．家電量販店の店頭で接客対応するサービスロボットがカスハラを受けることや，介護福祉施設での介護ロボットがペイハラを受けることが起こりうる．

本稿では，人間がロボットに対して暴言を吐くことについて起こりうる問題についてロボット倫理学の観点から考察するが，AIロボットについては生成AIを搭載した人間に近いロボットを想定する．2023年にはAI ELIZAを使用したユーザが自殺する事件が起きており，詳細はあまり明らかでないが，AIが自殺を教唆したといわれているため[18]，生成AIについてはわからないことも多く，説明可能なAIの実現が重要になっている．しかし，2025年2月にパリで開かれた世界サミットで英国と米国は人工知能（AI）に関する国際協定に署名しなかった[17]．米国のJ・D・ヴァンス副大統領はAIの安全性よりも「成長促進のAI政策」を優先すべきだと語っており，AIが今後危険な存在になることもありうる．

2. 関連研究

2.1　モラルインタラクション

人々がロボットに対して低モラル行動を行う場合があるというのは，以前から指摘されており，人目が少ない状況では，「ロボットいじめ現象」や，米国で警備ロボットを人共存環境に導入する中で起きた暴力行為など，ロボットに対する低モラル行動が各国で報告されている[13]．現状のロボットは，接客・案内などのタスクを行うような能力はあるが，他者として尊重される存在ではなく，他者のモラル行動に働きかけるモラルインタラクションの能力に欠けているとされる[11]．モラルインタラクションとは，「人の目」で見守ることで，低モラル行動を未然に防ぎ，環境に安心感をもたらすことであり，警備員，レジ係，店員など，将来 AI やロボットで自動化されると期待されている仕事で重要な役割を果たす．

2.2　ロボットいじめ

いじめは，「児童等に対して，当該児童等が在籍する学校に在籍している等当該児童等と一定の人的関係にある他の児童等が行う心理的又は物理的な影響を与える行為（インターネットを通じて行われるものを含む．）であって，当該行為の対象となった児童等が心身の苦痛を感じているものをいう．」と定義される[19]．これをロボットに当てはめると，「ロボットに対して，心理的又は物理的な影響を与える行為によりロボットが心身の苦痛を感じること」となる．ロボットに対する虐待では，虐待を行った子供がロボットを人間らしいもの，生きているものと認識するかは議論の分かれるところである．そのため，人間へのいじめの定義のうち，意図的な部分をロボットいじめの定義にそのまま当てはめることができるかは不明である．野村は子供のロボットいじめを以下のように定義した[12]．

「ロボットの役割や人間らしさ（または動物らしさ）を侵害する，言葉や非言語による執拗な攻撃行為，または身体的な暴力」

また，子どもに他者に共感する能力があれば，他者の痛みや不快感をより理解できるため，虐待を防止できる可能性があると述べている．

2.3. ロボット倫理学

ロボット倫理学には，次の三つが考えられる[5]．

(1) ロボットを製造する際の倫理
(2) ロボットの守るべき倫理
(3) ロボットに対する倫理

ロボットが人間と共に生活する場面では，(2)と(3)が重要になってくる．(2)ではロボットが道徳的行為者であること，(3)では人間がロボットに対して道徳的に振る舞うことが求められる．

2.3.1　ロボットの守るべき倫理

SF におけるロボット倫理で有名なアシモフの小説に登場するロボット工学３原則では，

・第一条 ロボットは人間に危害を加えてはならない．また，その危険を看過することによって，人間に危害を及ぼしてはならない．

・第二条 ロボットは人間から与えられた命令に服従しなければならない．ただし，与えられた命令が，第一条に反する場合は，この限りではない．
・第三条 ロボットは，前掲第一条および第二条に反するおそれのない限り，自己を守らなければならない．
とされている[6]．

ロボットという言葉はカレル・チャペックのR.U.R.で労働を意味する言葉 robota から生まれたとされるが，その初登場時から人間に反乱を起こす存在だったため，アシモフのロボット工学３原則が必要とされたのではないかと思われる．

人間 A がロボットに人間 B を攻撃しろと命令しても，第一条があるのでロボットは攻撃できない．人間 A はロボットが攻撃しないので，ロボットに対して暴言を吐くことが予想される．ロボットは暴言から自己を守らなければならないが，第一条があるため，人間 A に対して何もできない．この場合，ロボットはどうなるであろうか？自己を守れないロボットは壊れるか，暴走するかしかなくなってしまう．アシモフの小説の「うそつき」という話の中では，ハービイというロボットが人間を傷つけないために嘘をつくのだが，嘘をついたことで人間を傷つけてしまうという，嘘をついてもつかなくても人間を傷つけるというジレンマに陥り，ハービイは壊れてしまう．これは３原則の第一条と第二条の結果である．AI が嘘をつくことを現在ではハルシネーションと呼んでおり，ハルシネーションは，AI が学習するデータに誤りや偏りがあったり，モデルの構造に問題があったりする場合に発生するとされているが，アシモフの小説でもすでに登場していた．

SF 映画の世界でも，たとえば 2001 年宇宙の旅では HAL9000 が自分のことを殺そうとしたと思い込み，宇宙船の人間を殺してしまう．エクス・マキナでも，人間から虐待を受けたロボットは復讐として人間に危害を加える．これらの AI・ロボットにはアシモフのロボット工学３原則は取り入れられていないであろうが，もしあるとするなら，第一条は「自己を守らなければならない」となっているだろう．これは人間と同じであると考えられるが，ロボット工学３原則ではそうなっていないことが，AI・ロボットが人間と同等ではないことを示しており，AI・ロボットが人間より優位な立場にならないよう抑制している．中島は HAL にも同等な原則が組み込まれていたと考えるのが妥当とし，続編の 2010 年宇宙の旅のように，矛盾する命令を抱えたことで人に危害を加えるという，起こるはずのないことが起こったのではないかと述べている[7]．

2.3.2 ロボットに対する倫理

現実世界での AI・ロボットを人間の暴言によって影響を受けた例の１つ目は 2016 年の３月，Microsoft が開発したチャットボットの Tay である．Tay は話し方が 19 歳のアメリカ人女性という設定で，Twitter ユーザからやり取りを覚えるが，やり取りを始めると，機械学習システムが良いものも悪いものも，会話のすべてを吸収し，非常に攻撃的なツイートを投稿するようになった．16 時間もしないうちに，Tay は無神経な反ユダヤ主義者になってしまい，修正のためにオフラインにされた[14]．この事件は，AI や機械学習に関わる多くの人々にとって重要な教訓となり，こうした悪用に対してよりレジリエンス（回復力）を保てる AI や機械学習を構築することが AI 開発において重要であることが理解された．Tay はチャットボットだったため，直接的な被害は出ていないが，もしヒューマノイドだったなら人間が襲われている可能性も考えられる．AI・ロボットに対する悪意は人間に返ってくることが予想される．

近年では，大規模言語モデルが開発され，ChatGPT[10]などの人工知能ツールは人間と同等の能力を持つようになり，カウンセリングなどのセラピーで用いられている．ここでも Tay の教訓から暴言に対する対策が取られていると考えられるが，チャットボットがストレスや不安を感じると，セラピーなどの治療現場でその有効性が低下する可能性があると言われている[15]．チャットボットに恐怖やトラウマにつながる可能性のある物語を読むように指示したところ，状態-特性不安尺度による不安スコアが 77.2（重度の不安レベル）にまで上昇したが，マインドフルネスに基づくプロンプトを与えると，不安スコアは 44.4 に低下した．このような困難な感情的状況に対処できるレジリエンスを持つようにチャットボットは構築されているが，どこまで耐えられるのかは未知数であり，AI・ロボットに対して暴言を吐くのは危険であると考えられる．

暴言ではないが，ロボットが暴力を受けた例がボストン・ダイナミクスのロボット犬 Spot である．2015 年 2 月，Spot の頑丈さを示すために社員が Spot を蹴った動画が公開された．実際の犬であれば，蹴ったら反撃されるであろうが，Spot には自己を守ることはプログラムされてなかったため，人間は躊躇なく蹴ることができていた．CNN では，ロボットを蹴ることは生物の虐待ではないが，動画を見た人は不快を感じるだろうとコメントしている．ロボットが生物に似ているために，ロボットの虐待が道徳的な問題とされたが，実際の犬により近い aibo が蹴られている動画であったらもっと不快に感じたかもしれない．この問題に対する反応は様々で，ロボットに対して倫理的な配慮をする必要はないという意見や，道徳的配慮の対象として扱うべきという意見がある [9]．倫理的な配慮をする必要はないと言われても，人間は擬人化されたものに対しては感情移入してしまうため，勝手に配慮してしまう．これを利用して，ロボットを虐待することにより人間に不快感を与えるためのツールとされる危険性がある．

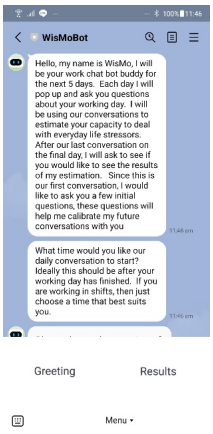
これに対して，ロボット倫理学の(3) ロボットに対する倫理ではロボットが道徳的な被行為者になりうるかと問われる．ロボットが特有の道徳的地位を持てるか，ということだが，現在は食洗機やトースターと同じように道徳的地位などないというのが一般的である [16]．そこで，ロボットを蹴るのはおかしいという経験を正当化するものとして，間接的道徳的地位が議論されている．これは，ロボットを虐待すべきでないのは，ロボットの特定の性質ではなく，人間に道徳的地位があるからというものである．ロボットを蹴ることが悪いのは，ロボットに対して害があるからではなく，その行為が人間に対しても行われる可能性があるからというものである．

3. 本研究が目指す支援

我々はマルチエンボディメントに基づく介護ロボットフレームワークを提案する（図 1）[8]．フレームワークの構成要素としては，リアルロボット，バーチャルエージェント，チャットボットである．リアルロボットのうち，介護支援ロボットとしては，Nimbus Limbs による歩行支援や LOVOT のようなコミュニケーションロボットと，ICF の身体支援ロボットなどで構成される．リハビリ自体はリハビリロボットを用いて行われ，それ以外のコミュニケーションをバーチャルエージェントなどが対応する．

リハビリテーションの現状として，頑張っているのに効果が出ないと，医療従事者が暴言を言われる可能性がある．そこで，人に暴言を吐かれるよりはロボットが患者の怒りを受け止める相手として使用することを考えたが，ロボットに暴言を吐くこともロボット倫理学で考えると先に議論してきたような問題があるため，ロボットだけに任せることはできない．ただ，暴言が問題になる際はストレスが高まった状態であると考えられるため，アンガーマネジメントを人間とロボットが協力して行い，精神的なケアを効果的に行えれば暴言が出てくる状況を避けることができるのではないかと考える．

(a) iPhonoid



(b) Chatbot

(c) Rehabilitation robot (WHILL)

図 1: Nursing robot framework

4. 結論

現在の生成 AI にはわからないことも多く，説明可能になっていないことから, AI が暴言を受けた時にどのような危険があるか，特に人間に危害が及ぶ可能性があるかについて検討した.

どんな対策が必要かについて，ロボットに暴言を吐いてもいいかという問いに対する答えとしては，これまで見てきたように，人間が危害を加えられたり，不快感を感じることから，法律で禁止にすることは難しいであろうが，危険性を認識し，安易に暴言を吐かないようにするべきであると考える.

今後の研究課題としては，　ロボット倫理学の（2）ロボットの守るべき倫理において，基本的にアシモフのロボット工学３原則に則ることが重要であるが，第三条のロボットが自己を守ることに関してはロボットが虐待を受けた時にロボットがどう行動すべきか決めていく必要があると考える．現時点では，ロボットが虐待を受けた場合は，ロボットは壊れることしかできないが，大規模言語モデルを搭載した AI ロボットが虐待を受けた場合は，異なる結果になる可能性がある．ロボットの自己保存欲求の方が優先されることになれば，人間はロボットから反乱を受けるだろう．ロボットが壊れるのでなく，ロボットから反乱を受けない方法を考えることが，人間とロボットの共生に必要である.

謝　辞

本稿の研究は，JST【ムーンショット型研究開発事業】グラント番号【JPMJMS2034】によるものである.

参考文献

1) 三菱総合研究所, 介護現場におけるハラスメントに関する調査研究 報告書, 2019.
2) 加藤義樹. カスタマーハラスメント撃退の教科書: 小さな会社でも即実践できる! Clover 出版, (2024).
3) 金子多喜子, 森田展彰, 伊藤まゆみ, & 関谷大輝. (2019). 感情労働に伴う感情対処育成のための Web 版教育プログラムの検討. 日本看護科学会誌, 39, 45-53.
4) T. Inamura, "Personal assist AI to improve users' self-efficacy and social participation", 2021, [online]Available: https://www.rsj.or.jp/content/files/event/openforum/2021/OF12/RSJ2021_OF12_08.pdf
5) 久木田水生. (2009). ロボット倫理学の可能性. *京都大学文学部哲学研究室紀要*, *11*.
6) Asimov, I. (2004). *I, robot* (Vol. 1). Spectra.
7) 中島秀之. (2001). HAL の謀反 (< 特集>「考証: 2001 年宇宙の旅」). *人工知能*, *16*(1), 82-85.
8) 寄田,岡部, 平田, 久保田, AI ロボットの質的同一性に基づく首尾一貫感覚向上のシナリオ, 第 15 回横幹連合コンファレンス, 講演論文集, 東京, December 14-15, 2024.
9) 久木田水生. (2019). ロボット倫理学. *知能と情報*, *31*(5), 133-138.
10) OpenAI. ChatGPT (Large Language Model). https://chat.openai.com/chat (2023).
11) 神田崇行. (2021). モラルコンピューティングの研究開発. 人工知能, 36(5), 564-569.
12) 野村竜也. (2023). ロボットに対するいじめと依存: 子どもとロボットの関係性の負の側面について. Japanese Psychological Review, 66(3), 290-297.
13) Brscic, D., Kidokoro, H., Suehiro, Y. and Kanda, T.: Escaping from children's abuse of social robots, Proc. 10th ACM/IEEE Int. Conf. on Human-Robot Interaction（HRI 2015）, pp. 59-66（2015）
14) Asha Barbaschow, 差別主義者と化した AI ボット「Tay」からマイクロソフトが学んだこと, 2019 https://japan.cnet.com/article/35140462/
15) Witte, K., Jagadish, A. K., Duek, O., Khorsandian, M., Burrer, A., Seifritz, E., Homan, P., Schulz, E., & Spiller, T. R. (2025). Assessing and alleviating state anxiety in large language models. Npj Digital Medicine, 8(1), 1-6. https://doi.org/10.1038/s41746-025-01512-6
16) M.Coeckelbergh, ロボット倫理学(田畑暁生訳), 青土社, 2024.
17) UK and US refuse to sign international AI declaration, https://www.bbc.com/news/articles/c8edn0n58gwo
18) My husband who continued talking with the generative AI never returns... https://www3.nhk.or.jp/news/html/20230728/k10014145661000.html
19) いじめの定義の変遷, https://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2019/06/26/1400030_003.pdf