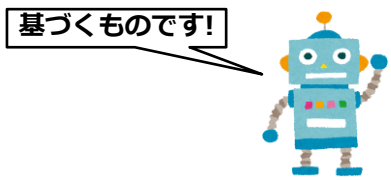
ロボットは謝るべきか?
―クリティカル・ロボティクス 序説―

呉羽 真
山口大学 国際総合科学部
講師 / 博士 (文学)

1

謝辞

本研究は、日本学術振興会科学研究費補助金・新学術領域研究 (研究領域提案型) **の課題「対話知能システムの研究開発及び社会実装のための法社会規範の研究」**
(研究代表者: 新保史生, 領域代表者: 石黒浩,
課題番号: 19H05694)
に基づくものです。



2

アウトライン

1. 背景と問題: 謝るロボット
2. 謝罪とは何か?
3. ロボットは謝るべきか?
4. 問題提起するロボット
   ―クリティカル・ロボティクスの挑戦―

3

1. 背景と問題: 謝るロボット

4

背景: 謝るロボット

よくある考え方:
(ロ)ボットは謝らない or うまく謝れない?

- 例1) Tayのヘイトスピーチ事件 (2016)
  ➡ 謝罪したのはMicrosoft
- 例2) 山田胡瓜『AIの遺電子』
- 例3) 秋本治『こちら葛飾区亀有公園前派出所』

しかし近年、**謝る(ロ)ボットが登場!**

- 例4) エリカ「すみません、うまく聴き取れなかったので、もう1度言っていただけますか?」(Uchida et al. 2019)
- 例5) Amazon EchoのAlexaは、謝罪を要求されるといつも従うようプログラムされている (Strengers 2021)
- 例6) ChatGPTもよく謝る

5

フィクションに描かれた謝罪ロボット: 『AIの遺電子』より

「第28話 謝罪」(山田 2016)

- クレーム対応をするロボット (ヒューマノイド) のエピソード
- クレーマーが謝罪に来たロボットに言った台詞:
  「ふざけんじゃねぇ!! プログラムに謝られたってな! なんの誠意にもなってねーんだよ!! 人間が出張ってこい!!」(109～110頁)
- ストレスで悩むロボットのために、会社はクレーム対応を外部の人間 (プロ) に委託する。だが、彼らが見せた対応は、非人間的なものだった…

6

フィクションに描かれた謝罪ロボット:『こちら葛飾区亀有公園前派出所』より

「ロボット時代の巻」(秋本 2015)

- 両津が謝罪ロボットを含む様々なソーシャルロボットを開発･販売するエピソード
  - 両津:**「サラリーマン型ロボットはこれから売れるぞ ニュースで謝るシーンあるだろ 絶対に『なんで謝る必要あるんだよ』と思っているそれならいっそロボットでいいだろ」**(253～254頁)、**「どうせパフォーマンス的だし」**(254頁)
- 海外で接待ロボが人気になるが、突然笑い出す誤作動を起こし、クレームが寄せられる
- 両津は記者会見で自分の代わりに「申し訳ないロボ」に謝罪させるが、非難殺到
  - 麗子:**「誠意がまったく無いわね」**(263頁)

7

ロボット謝罪の効果

HRIの知見:

- ロボット謝罪には、**信頼回復効果** (Lee et al., 2010; Robinette et al., 2015; de Visser et al. 2018; Kim & Song 2021; Fratczak et al., 2021; Pompe et al., 2022) がある
- **対話破綻回避効果**も? (Uchida et al. 2019)

8

ロボット謝罪の意義

ロボット謝罪は、人間とのレジリエントな関係の構築を可能にする?

- どれほどロボット技術が進歩しても、人間に不利益を及ぼす可能性はゼロにはならない
- これは人間も同じ。だが人間は、他人に不利益を及ぼしたとき、**謝ったり赦したりする**ことで、社会を維持している
- ロボットは人間と、このレジリエントな関係を築けるか?

ロボット謝罪は、人間を感情労働から解放する?

- クレーム対応のような仕事は、情動の制御を求められ、精神的消耗を伴う**「感情労働」**
- ロボットにやらせればよい? (e.g. Kim 2017)

9

問題

疑問: **何でエリカが謝ってるの???**

- ロボットが人間を模した行動をとることは、倫理学で問題視されてきた (e.g. Sparrow & Sparrow 2006; Turkle 2011)
- ロボット謝罪は、社会に良くない影響をもたらすかも

⬇

問題:
ロボットは (どんな場合に)
謝るべき/謝ってよいか?

10

2.謝罪とは何か?

11

謝罪研究の広がり

- **言語哲学**
  - 例) Austin (1962) とSearle (1969, 1979): 謝罪を言語行為として分析
- **倫理学・政治哲学・社会哲学**
  - 例) Smith (2008): 謝罪の機能と意味を分析
- **社会学**と**言語学**
  - 例) Goffman (1971): 謝罪を「面子 face」を繕うための「修復作業」として分析
    ➡ Brown & Levinson (1987) らの「ポライトネス理論」に継承
- **心理学**
  - 例) 大渕 (2010): 謝罪を釈明の一種と見なし、その行い手と受け手の心理、文化差を調査

12

謝罪とは何か
―釈明の一種としての謝罪―

「釈明 account」の種類 (e.g. 大渕 2010)

- **「謝罪 apology」**: **(a)自分が当該の行為をしたこと、(b)その行為が間違っていること、(c)責任が自分にあること**、を認める
  - 例)「私がやりました。すみません」
- **「弁解 excuse」**: (c)を認めず、責任が自分にないと主張する
  - 例)「やったのは私ですが、○○さんにそうしろと言われました」
- **「正当化 justification」**: (b)を認めず、行為が正しかったと主張する
  - 例)「確かにやりましたが、それには～という理由があります」
- **「否認 denial」**: (a)を認めず、自分はその行為をしていないと主張する
  - 例)「私じゃないです。どうせ○○さん辺りでしょう」

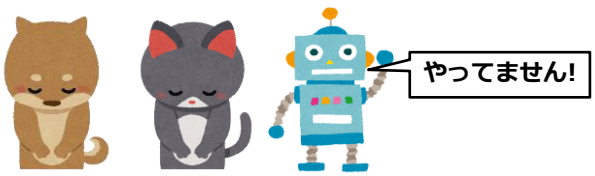
13

謝罪とは何か
―謝罪の要素と目的―

謝罪に求められる要素 (Gill 2000; Smith 2008; 大渕 2010; Radzik & Murphy 2021)

- 間違った行為 (過失/不正。不作為を含む) の事実の承認
- 行為に対する自身の責任の承認
- 行為のどこが間違っていたかの認識の表明
- 被害者へのいたわり/尊重の表明
- 後悔の気持ちの表明
- 改善の意志や方法の表明
- など

目的:
- 和解/赦しの達成
- 信頼の回復



14

謝罪とは何でないか
―non-apology apology―

「謝罪もどき non-apology apology」…
謝罪に求められる要素を欠く、擬似謝罪行為
(望月 2021; cf. 古田 2021)

- 例1)**「遺憾に思います」**
  ➡ 間違った行為が行われたことを認めるだけで、行為に対する責任を負わず、どこが間違っていたかも特定しない
- 例2)**「不快に思った方がいたならば申し訳ございません」**
- 例3)**「誤解を与えて申し訳ございません」**
- 例4)**「お騒がせして申し訳ございません」**
  ➡ 行為の間違いを被害者や世間の受け取り方の問題にすり替え、間違った行為をした事実を曖昧にする

15

3.ロボットは謝るべきか?

16

ロボットは謝るべきか?
—デフォルト的反対論—

デフォルト的反対論: **ロボットは謝るべきでない**

- 謝罪は本質的に、自身の責任の承認を伴う (e.g. 大渕 2010)。だが、現状のロボットは自らの振る舞いに責任を負いえない
- 責任を負うのはロボットの背後にいる人間
  ➡ ロボットが謝ることで、人間の責任が曖昧になる
  ➡ 真の和解/赦しを妨げる
  - 例) ロボットがTayと同様のヘイトスピーチを行った場合、ロボットが謝ることで、開発者やその所属組織の責任が曖昧になりかねない
  - HRIの知見: 人々はロボットに対し、その振る舞いに応じて、一定の責任を帰属する (Kahn et al. 2012; Shank et al. 2019; 河合 2020)
- ロボットによる謝罪が頻繁に行われるようになれば、人間による真の謝罪が重みを失うかもしれない???

17

ロボットは謝るべきか?
—デフォルト的反対論—

デフォルト的反対論: **ロボットは謝るべきでない**

- 謝罪は本質的に、自身の責任の承認を伴う (e.g. 大渕 2010)。だが、現状のロボットは自らの振る舞いに責任を負いえない
- 責任を負うのはロボットの背後にいる人間
  ➡ ロボットが謝ることで、人間の責任が曖昧になる
  ➡ 真の和解/赦しを妨げる
  - 例) ロボットがTayと同様のヘイトスピーチを行った場合、ロボットが謝ることで、開発者やその所属組織の責任が曖昧になりかねない

発表者の見解:
この反対論が当てはまるケースがある。
が、ロボット謝罪一般に当てはまるわけではない

- 深刻な害のないケースもある
- 人間だって、謝らなくていい場面で謝りまくっている

18

補足:
ロボットの謝罪表現の使用は儀礼的?

疑問: **ロボットの「すみません」は本当に謝罪?**

- 謝罪表現は**儀礼的**に用いられる
  - 例)「すみません、注文 (or 質問) いいですか?」、「すみません、忘れ物ですよ」
    ➡ 本気で過失や不正、その責任を認めているわけではなく、真の謝罪ではない
- これはエリカの「すみません」には当てはまらない
  - Uchida et al. (2019) の見解: 謝罪は (非難と並んで) ロボット-ユーザー間で対話破綻の責任を分担する手段。まずロボットが謝り、次に非難することで、ユーザーの協力意図を引き出し、対話破綻を回避できる
    ➡ 軽度であれ、過失と責任の承認を行っている

19

想定される反論:
フィクション的解釈

想定される反論: **人々は本気でロボットに責任がある (=非難に値する) と思っていないのでは?**

- **フィクション的解釈**: 人々のロボットへの態度は、フィクションに対する反応になぞらえて理解できる?
  (e.g. Sharkey & Sharkey 2006; Duffy et al. 2012; Rodogno 2016; 久木田 2017; 水上 2020)
  - ユーザーは、本心からロボットに感情があると思っているわけではなく、**「不信の宙吊り」**or**「ごっこ遊び」**に興じているだけ
  - 経験的根拠: 人々は、ロボットが感情をもつかのように振る舞うが、明示的にロボットに感情があるか聞かれると否定する (Sharkey & Sharkey 2006; Gray & Wegner 2012)

⬇ **責任帰属にも適用**

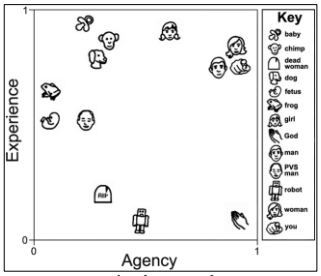
- ユーザーは、ロボットの謝罪を受け入れてしまうときでも、本心からロボットに責任があると思っているわけではない?

20

想定される反論への応答①: 責任帰属に関する経験的知見

想定される反論への応答①:
フィクション的解釈は、ロボットへの責任帰属にそのまま当てはめることはできない

- **HRIの知見**: 人々はロボットに対し、その振る舞いに応じて、明示的に一定の責任を帰属する (Kahn et al. 2012; Shank et al. 2019; 河合 2020; 谷部 2020)
- **社会心理学の知見** (Gray et al. 2007)
  - 心の帰属は、「行為者性 agency」と「経験 experience」の2次元
  - 責任に関わる行為者性は、感情に関わる経験と違い、ロボットにもある程度帰属される



Gray et al. (2007), p. 619

21

想定される反論への応答②: 現実への影響

想定される反論への応答②: **ロボットの謝罪は、現実に人間の行動に影響を与える可能性がある**

①ロボットへの態度とフィクションへの反応には違いがある
- 私たちは、実世界で作動しているロボットと、能動的にやり取りする ➡ フィクション作品の鑑賞と比べて、現実/虚構の境界が不鮮明になりやすい

②フィクションだからといって何でも許されるわけではない
- 例) (Tayが行ったような) ヘイトスピーチは、害が大きいため、フィクションの中でも安易に登場人物に言わせてはならない
  ➡ 深刻な害が生じうる行為に関しては、ロボットの謝罪は、フィクション的解釈によって正当化されない
- フィクションにおける描写が悪影響をもたらす可能性もある? (坂元編 2011; Dill-Shackleford 2015)

22

ロボット謝罪がもたらしうる悪影響の再考

①ロボット謝罪が誘発する責任の誤帰属が、真の和解や赦し、改善を妨げる

②ロボット謝罪が頻繁に行われれば、人間による真の謝罪が重みを失うかもしれない?

- 深刻な過失や不正の場合は、人が謝るべき
  - 例) (Tayが行ったような) ヘイトスピーチ
- ただし、深刻な過失や不正がない場合もある
  - 例) エリカのケース。問題になっている過失は、音声認識の失敗という、些細なもの
    ➡ このケースが和解/赦しや改善を要求するわけではない

23

ロボット謝罪がもたらしうる悪影響の再考

①ロボット謝罪が誘発する責任の誤帰属が、真の和解や赦し、改善を妨げる

②ロボット謝罪が頻繁に行われれば、人間による真の謝罪が重みを失うかもしれない?

反論: **これは人間の謝罪にも当てはまる**

- 日本社会に蔓延する「謝罪もどき」
- 儀礼的謝罪表現の使用
  - 例)「すみません、注文 (or 質問) いいですか?」
- 謝らなくていい場面での謝罪
  - 例) 2020北京五輪での高梨沙羅選手の謝罪
- 「謝罪代行業」

24

ロボットは謝るべきか?
―ここまでのまとめと結論―

まとめ: デフォルト的反対論が当てはまるケースがある。
が、**ロボット謝罪一般に当てはまるわけではない**

①責任の誤帰属が深刻な害をもたらさないケースもある

②人間だって、謝らなくていい (自分に責任がない) 場面で謝りまくっている

結論: **ロボット謝罪が許されるかは、ケースバイケースで判断するしかない**

- 深刻な害のないケースでは、謝罪を巡る人間社会の慣行に照らせば、ロボットに謝罪してはならないとは言えない

25

ロボットは謝るべきか?
―ここまでのまとめと結論―

さらなる疑問: **この慣行自体に問題があるのに、それを踏襲していいのか?**

- 深刻な害のないケースでは、そもそも人間も謝る必要はない
- 軽度の過失に対して過剰な対応を要求する悪しき文化を、ロボット謝罪は助長してしまうのでは?

- 深刻な害のないケースでは、**謝罪を巡る人間社会の慣行**に照らせば、ロボットに謝罪してはならないとは言えない

呉羽のアイディア:
ロボットを用いてこの慣行自体を問い直せないか?

26

4.問題提起するロボット —クリティカル・ロボティクスの挑戦—

27

クリティカル・ロボティクス

従来のロボティクスの目的 (e.g. 石黒 2021)：
ロボットの開発を通して…

①技術的側面: **社会の役に立つものを作る**
- 社会に受け入れられた価値観を前提

②科学的側面: (構成論的アプローチを通して) **人間を理解する**
- 価値中立的 (?)

＋

③社会に受け入れられた価値観に対して問題を提起する
➡「クリティカル・ロボティクス」
- 「クリティカル・デザイン」をロボットに適用
- モデル:「役に立たない機械」,「弱いロボット」

28

クリティカル・デザイン

「クリティカル・デザイン」
(Dunne & Raby 2013, ch. 3)
技術の設計を通して、社会に受け入れられた先入観や固定観念に揺さぶりをかけることを目指すデザインのアプローチ

- 技術哲学でも言及あり (Michelfelder 2017)

ロボットの設計に適用

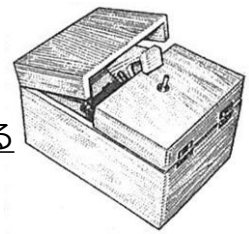
- ロボットは、人間のあり方を映し出す**「鏡」**(石黒 2009) としての象徴的意味をもつ ➡ 批判的機能を豊かにもつ
  - 「ロボットのふり見てわがふり直せ」

29

クリティカル・ロボティクスの例

「役に立たない機械 useless machine」
(Minskyが考案; cf. Pesta 2013)

- スイッチをオンにされるとオフにするだけ
- 「機械は役に立たなければならない」と考える社会への問題提起



Pesta (2013) より転載

「弱いロボット」(岡田 2012)

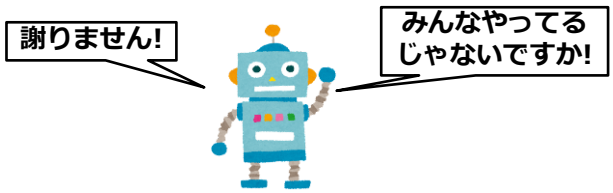
- 例) 自分でゴミを拾えないゴミ箱ロボット
  - 落ちているものを見つけるとそれを周囲の人に教える仕草をする。それを見た人はゴミを拾ってあげたくなる
- 「1人でできる強さ」を重視する社会への問題提起

30

謝罪に関するクリティカル・ロボティクス

疑問: **人間社会の謝罪慣行に対して問題提起を行うロボットは、具体的にどんなものか?**

- 例1) 軽度の過失に対して過剰な対応を要求する文化を、**「謝らない (謝ったら死ぬ) ロボット」**で批判?
- 例2) 謝罪もどきで責任逃れする文化を、**「露骨に責任逃れするロボット」**で批判?



31

文献①

- Austin, J.L. 1962. *How to Do Things with Words*, Clarendon Press. [2019.『言語と行為』飯野勝巳訳, 講談社]
- Brown, P. & Levinson, S.C. 1987. *Politeness*, Cambridge University Press. [2011.『ポライトネス』田中典子監訳, 研究社]
- de Visser, E.J., et al. 2018. 'From 'automation' to 'autonomy,' *Ergonomics* 61: 1409-1427.
- Dill-Shackleford, K.E. 2015. *How Fantasy Becomes Reality (Revised and Expanded Edition)*, Oxford University Press. [2019.『フィクションが現実となるとき』川端美樹訳, 誠信書房]
- Duffy, B.R. & Zawieska, K. 2012. 'Suspension of disbelief in social robotics', *2012 IEEE RO-MAN*, pp. 484-489.
- Dunne, A. & Raby, F. 2013. *Speculative Everything*, MIT Press. [2015.『スペキュラティヴ・デザイン』千葉敏生訳, ビー・エヌ・エヌ新社]
- Fratczak, P., et al. 2021. 'Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction,' *International Journal of Industrial Ergonomics* 82, March 2021, 103078.
- Gill, K. 2000. 'The moral functions of an apology', *Philosophical Forum* 31: 11-27.
- Goffman, E. 1971. *Relations in Public*, Routledge.
- Gray, H.M., Gray, K., & Wegner, D.M. 2007. 'Dimensions of mind perception', *Science* 315: 619.

32

文献②

- Gray, K. & Wegner, D.M. 2012. 'Feeling robots and human zombies', *Cognition* 125: 125-130.
- Kahn, P.H., et al. 2012. 'Do people hold a humanoid robot morally accountable for the harm it causes?', *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 33-40.
- Kim, M. 2017. 'Let robots handle your emotional burnout at work', *How We Get To Next*, March 23, 2017. URL=<https://www.howwegettonext.com/let-robots-handle-your-emotional-burnout-at-work/>
- Kim, T. & Song, H. 2021. 'How should intelligent agents apologize to restore trust?' *Telematics and Informatics* 61: 101595
- Lee, M.K., et al. 2010. 'Gracefully mitigating breakdowns in robotic services,' *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 203-210.
- Michelfelder, D., et al. 2017. 'Designing differently,' S.O. Hansson (ed.), *The Ethics of Technology*, pp. 193-218, Rowman & Littlefield.
- Pesta, A. 2013. 'Looking for something useful to do with your time? don't try this,' *Wall Street Journal*, March 13, 2013: 1&A12.
- Pompe, B.L., et al. 2022. 'The Robot that showed remorse,' *2022 IEEE RO-MAN*, pp. 260-265.
- Radzik, L. & Murphy, C. 2021. 'Reconciliation', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. URL=<https://plato.stanford.edu/entries/reconciliation/>

33

文献③

- Robinette, P., et al. 2015. 'Timing is key for robot trust repair', *Social Robotics. ICSR 2015. Lecture Notes in Computer Science*, vol. 9388. Springer.
- Rodogno, R. 2016. 'Social robots, fiction, and sentimentality', *Ethics of Information Technology* 18: 257-268.
- Searle, J.R. 1969. *Speech Acts*, Cambridge University Press. [1986.『言語行為』坂本百大·土屋俊訳, 勁草書房]
- ――― 1979. *Expression and Meaning*, Cambridge University Press. [2006.『表現と意味』山田友幸訳, 誠信書房]
- Shank, D.B., et al. 2019. 'When are artificial intelligence versus human agents faulted for wrongdoing?' *Information, Communication & Society* 22: 648-663.
- Smith, N. 2008. *I Was Wrong*, Cambridge University Press.
- Sparrow, R. & Sparrow, L. 2006. 'In the hands of machines?', *Mind & Machines* 16: 141-161
- Turkle, S. 2011. *Alone Together*, Basic Books. [2018.『つながっているのに孤独』渡会圭子訳, ダイヤモンド社]
- Strengers, Y. 2021. 'Amazon Echo's Alexa is programmed to always apologize — especially when it's not her fault', *NBC News Think*, March 2, 2021. URL=<https://www.nbcnews.com/think/opinion/amazon-echo-s-alexa-programmed-always-apologize-especially-when-it-ncna1259001>

34

文献④

- Uchida, T., et al. 2019. 'Who is responsible for a dialogue breakdown?' *Frontiers in Robotics & AI* 6: 29.
- 秋本治 2015.『こちら葛飾区亀有公園前派出所 195』, 集英社.
- 石黒浩 2009.『ロボットとは何か』, 講談社.
- ――― 2021.『ロボットと人間』, 岩波書店.
- 大渕憲一 2010.『謝罪の研究』, 東北大学出版会.
- 岡田美智男 2012.『弱いロボット』, 医学書院.
- 河合祐司 2020.「ロボットへの原因と責任の帰属」,『日本ロボット学会誌』38: 32-36.
- 久木田水生 2017.「AIと誠」, 久木田水生·神崎宣次·佐々木拓『ロボットからの倫理学入門』所収, 105～118頁, 名古屋大学出版会.
- 坂元章編 2011.『メディアとパーソナリティ』ナカニシヤ出版.
- 谷辺哲史 2020.「AI·ロボット工学と社会性認知」, 唐沢かおり編『社会性認知』所収, 151～165頁, ナカニシヤ出版.
- 古田徹也 2021.『いつもの言葉を哲学する』, 朝日新聞出版.
- 水上拓哉 2020.「ソーシャルロボットの倫理のための概念工学」,『2020年度人工知能学会全国大会 (第34回) 論文集』.
- 望月優大 2021.「「謝らない謝罪」が日本で蔓延している」,『ニューズウィーク日本版』2021/7/28. URL= <https://www.newsweekjapan.jp/mochizuki/2021/07/post-8_2.php>
- 山田胡瓜 2016.『AIの遺電子 3』, 秋田書店.

35