ロボットは謝るべきか?

呉羽 真
山口大学 国際総合科学部 講師
博士 (文学)

1

ロボットと謝罪

(ロ)ボットはふつう謝らない

- 例) Tayのヘイトスピーチ (2016) ➡ 謝罪したのはMicrosoft

ロボットはうまく謝れない

- 例) 山田胡瓜『AIの遺電子』「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!!」(山田 2016, 109～110頁)

謝るロボットもいる!

- 例) 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)
- 謝罪は、人間との円滑な相互作用･関係構築を可能にする
  - 信頼回復効果 (Robinette et al. 2015; de Visser et al. 2018; Kim & Song 2021)
  - 対話破綻回避効果 (Uchida et al. 2019)

2

問題意識: なぜ謝罪を取り上げるのか

謝罪は社会を維持するための重要な手段

- 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?

謝罪はロボットの有望な用途?

- 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)

だが、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)

3

問題

疑問:
ロボットは (どんな場合に)
謝るべき/謝ってよいか?

⬇

「人間機械共生社会」におけるロボットの位置づけを考えるための試金石

4

1. 謝罪とは何か

5

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

6

謝罪とは何か
―釈明の一種としての謝罪―

「釈明 account」の種類 (大渕 2010)

- **「謝罪 apology」**: (a)自分が間違った行為をしたこと、および、(b)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」**: (b)を認めず、責任が他者にあると主張する
  - 例)「やったのは私ですが、K山さんにそうしろと言われました」
- **「正当化 justification」**: (a)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K山さん辺りでしょう」

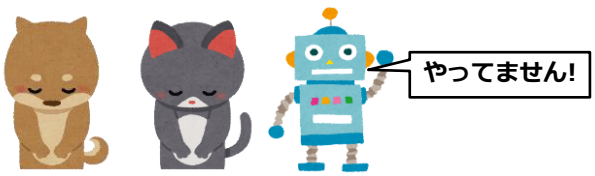
7

謝罪とは何か
―謝罪の要件と目的―

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



8

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

- 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」**
  ➡ 行為の間違いを被害者の受け取り方の問題にすり替え、行為が本当に間違っていたかどうかを曖昧にする
- 例4) **「お騒がせして申し訳ございません」**
  ➡ 世間を騒がせたことを認めるだけで、間違った行為をしたことを曖昧にする

9

2. ロボットは謝るべきでない?
―デフォルト的反対論とその検討―

10

ロボットは謝るべきでない?
―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、自身の責任の承認を伴う。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*
  - 例) Tayのヘイトスピーチ: MicrosoftのTayチームには、荒らし対策を十分に行わなかった、という過失がある (Jeong 2016)
    ➡ ロボットが同様の発話を行った場合、ロボットが謝ることで、開発者や彼らが属する組織の責任が曖昧になりかねない
- 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
- * ロボットが謝ることは、人間が謝らないことを含意しない。が、結局人間が謝るならロボットに謝らせる意味は不明

11

ロボットは謝るべきでない?
―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、自身の責任の承認を伴う。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*
  - 例) Tayのヘイトスピーチ: MicrosoftのTayチームには、荒らし対策を十分に行わなかった、という過失がある (Jeong 2016)
    ➡ ロボットが同様の発話を行った場合、ロボットが謝ることで、開発者や彼らが属する組織の責任が曖昧になりかねない
- 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
- * ロボットが謝ることは、人間が謝らないことを含意しない。が、結局人間が謝るならロボットに謝らせる意味は不明

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
- 人間だって、謝らなくていい場面で謝りまくっている
- 人々がそれほど人間とロボットを混同するかは疑問

12

ロボット謝罪の是非を左右する要因

①ロボットの用いる謝罪表現が、本当に謝罪のためのものか、それとも単なる儀礼的なものか?

②ロボットの背後にいる人間に実際どれだけの過失や不正があるか?

③人々がどれだけ本気でロボットに責任があると思っているか?

④ロボットの謝罪が、人間の行動にどれだけの影響を与えるか?

13

謝罪表現の儀礼的性格

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

14

過失/不正の深刻さ

疑問: **ロボットの背後にいる人間に実際どれだけの過失 (や不正) があるか?**

- 責任の誤帰属の害: 真の和解/赦し、改善を妨げる
- エリカの事例では、深刻な過失や不正はない
  - 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。同じ発話を繰り返す手間をかけるだけ
  - このケースが和解/赦しや改善を要求するわけではない
- より深刻な過失や、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
  - 疑問: これらの場合、人々はロボットの謝罪を受け入れるのか?

15

責任帰属の真正性

疑問: **人々はどれだけ本気でロボットに責任がある (=非難に値する) と思っているか?**

1つの解釈: 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)

- 例) エリカが「すみません」と言うとき、ユーザーは、無自覚的・自動的にエリカの謝罪を受け入れてしまうとしても、本心からエリカに責任があると思っているわけではない

⬇

- この解釈の責任帰属への適用は、現在得られている経験的知見からは、支持されない。が、より詳細な調査に値する

16

ロボットへの態度のフィクション的解釈 ―感情帰属―

ロボット倫理学の問題: **感情をもつかのように振る舞うロボットの開発は、ユーザーに対する欺瞞?**
(Sparrow & Sparrow 2006)

応答: **人々は騙されていない**

- HRI・社会心理学の知見: 人々は (無自覚的に) ロボットが感情をもっているかのように振る舞う。だが、明示的にロボットが感情をもっているか問われると、否定する人が多い (Sharkey & Sharkey 2006; Gray & Wegner 2012)
- 解釈: **「不信の宙吊り」**(Sharkey & Sharkey 2006; Duffy et al. 2012) or **「ごっこ遊び」**(Rodogno 2016; 久木田 2017) に興じ、感情をもたないロボットとのやり取りを楽しんでいる

17

ロボットへの態度のフィクション的解釈 ―感情帰属―

元ネタ: **フィクションの美学**

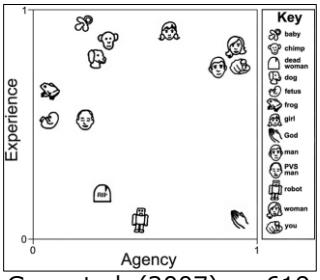
- **問題** (「フィクションのパラドックス」): ホラー映画に登場する幽霊やゾンビに対して恐怖を抱くのは不合理に思われる
- **応答**: 人々は、(1)幽霊やゾンビはいないという**信念を「宙づり」**して怖がっている (Coleridge 1907) or (2)あたかも怖がっているかのような**「ごっこ遊び」**をしている (Walton 1978)

- 解釈: **「不信の宙吊り」**(Sharkey & Sharkey 2006; Duffy et al. 2012) or **「ごっこ遊び」**(Rodogno 2016; 久木田 2017) に興じ、感情をもたないロボットとのやり取りを楽しんでいる
- 疑問: ロボットへの感情帰属に関するフィクション的解釈を、責任帰属にも拡張可能? (水上 2020)

18

ロボットへの態度のフィクション的解釈 —責任帰属に関する経験的知見—

- **HRIの知見**: 人々はロボットに対し、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)
- **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
    ➡ 道徳的行為者性と受容者性に対応
  - 謝罪の問題で問われる「行為者性」は、欺瞞の問題で問われる「経験」と違い、ロボットにもある程度帰属される
  - だが、ロボットの行為者性は、真の行為者性とは区別されているかも? (谷辺 2020)
    ➡ より詳細な調査が必要



Gray et al. (2007), p. 619

19

影響の大きさ —フィクション的解釈の限界—

疑問: **ロボットの謝罪は、人間の行動にどれだけの影響を与えるか?**

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 (暴力や差別の助長) をもたらす可能性もある? (坂元編 2011; Dill-Shackleford 2015)

20

影響の大きさ
—フィクション的解釈の限界—

疑問: **ロボットの謝罪は、人間の行動に<u>どれだけの影響</u>を与えるか?**

ありうる懸念: **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**

- 例)
  - 日本社会に蔓延する「謝罪もどき」(望月 2021; 古田 2021)
  - 儀礼的謝罪表現の使用
  - 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)
- 文化差もある (大渕 2010)

21

3. 結論
—ロボットは謝るべきか?—

22

結論

ロボット謝罪の是非は、今後の経験的研究の進展を見守りつつ、ケースバイケースで判断する必要がある

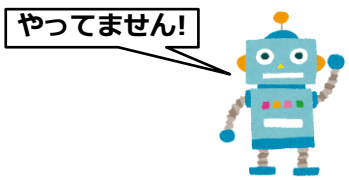
- 現実に生じているケース (例: エリカの事例) は、深刻な過失や不正がなく、大きな害が生じない、という理由で、取り立てて問題はなさそう
- だが、深刻な過失/不正に関する責任の誤帰属と、大きな害をもたらすがゆえに、避けられるべき場合もありうる

- 経験的に明らかにされるべきこと:
  - 深刻な過失や不正に関しても、人々はロボットの謝罪を受け入れるか?
  - 人々がどれだけ本気でロボットに責任があると思っているか?
  - ロボットの謝罪が、人間の行動にどれだけの影響を与えるか?
  - ロボットの謝罪への反応に文化差はあるか?

23

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。



24

文献①

- Austin 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.]
- Coleridge, S.T. 1907. *Biographia Literaria*, Oxford University Press. [1976.『文学評伝』桂田利吉訳, 法政大学出版局.]
- de Visser, E.J., Pak, R. & Shaw, T.H. 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.
- 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.
- Gray, K. & Wegner, D.M. 2012. 'Feeling robots and human zombies', *Cognition* 125: 125-130.

25

文献②

- Jeong, S. 2016. 'How to make a bot that isn't racist', *Vice*, March 26, 2016. URL = <https://www.vice.com/en/article/mg7g3y/how-to-make-a-not-racist-bot>
- Kahn, P.H., Kanda, T., Ishiguro, H., Gill, B.T., Ruckert, J.H., Shen, S., Gary, H.E., Reichert, A.L. Freier, N.G. & Severson, R.L. 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
- 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/>
- Robinette, P. Howard, A.M. & Wagner, A.R. 2015. 'Timing is key for robot trust repair', in A. Tapus, E. André, J.C. Martin, F. Ferland & M. Ammi eds., *Social Robotics*, Springer.

26

文献③

- 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., DeSanti, A. & Maninger, T. 2019. 'When are artificial intelligence versus human agents faulted for wrongdoing?', *Information, Communication & Society* 22: 648-663.
- Sharkey, N. & Sharkey, A. 2006. 'Artificial intelligence and natural magic', *Artificial Intelligence Review* 25: 9-19.
- 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.]
- Uchida, T., Minato, T., Koyama, T. & Ishiguro, H. 2019. 'Who is responsible for a dialogue breakdown?', *Frontiers in Robotics & AI* 6: 29.
- Walton, K. 1978. 'Fearing fictions', *Journal of Philosophy* 75: 5-27. [「フィクションを怖がる」森功次訳, 西村清和編·監訳『分析美学基本論文集』所収, 301～334頁, 勁草書房.]

27

文献④

- 大渕憲一 2010.『謝罪の研究』東北大学出版会.
- 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(1): 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.

28