謝罪を行う対話ロボットの評価に関する一考察

酒造 正樹 [1]　安部 健太 [2]　今井 久登 [3] 久保山 哲二 [3]　磯上 貞雄 [3]
[1] 湘南工科大学 [2] 帝京大学 [3] 学習院大学
shuzo@info.shonan-it.ac.jp　kenta.abe@main.teikyo-u.ac.jp
hisato.imai@gakushuin.ac.jp　ori-nlp2026@tk.cc.gakushuin.ac.jp
isogami@gakushuin.ac.jp

概要

謝罪機能を備えた対話ロボットや対話エージェントは，失敗回復や関係維持を目的として研究が進められてきた。しかし，謝罪は単なる発話生成ではなく，責任の所在，感情への配慮，関係修復への姿勢を含む社会的行為であり，その評価は文脈や受け手の解釈に強く依存する。本研究では，謝罪を行う対話主体に対する主観評価実験の経験にもとづき，従来の評価尺度では謝罪行動の差異や評価者の判断過程を十分に捉えきれない事例が生じることを整理する。評価が困難となる状況を事例として示すことで，謝罪行動の評価において考慮すべき前提条件を明確にし，今後の評価研究や評価尺度構成に向けた検討の基礎を与えることを目的とする。

1　はじめに

近年，対話ロボットや対話エージェントには，情報提供や案内といった機能に加えて，利用者との関係を円滑に維持するための社会的振る舞いが求められている。とくに，誤案内や処理失敗といった不具合が生じた場面において，謝罪を行う機能を備えた対話主体が提案されており，失敗回復や信頼回復の観点から一定の効果が報告されている [1, 2]。

一方で，謝罪は単なる発話生成ではなく，責任の所在，感情への配慮，関係修復を含む社会的行為であり，その成否は文脈や受け手の認知に強く依存する。このため，謝罪を行う対話ロボットをどのような観点で評価すべきかについては，いまだ整理が十分になされていない。多くの先行研究では，満足度や印象評価といった主観評価尺度を用いた実験が行われているが，謝罪行動の差異がこれらの指標に明確に反映されない事例も報告されている。

我々はこれまで，過失を犯した学生役を想定した三次元 CG アバター型の謝罪対話ロボットを複数構築し，言語表現，音声的特徴，表情変化を統合したマルチモーダルな謝罪行動を実装してきた [3]. しかし，従来型の主観評価指標を用いた実験では，謝罪行動の設計差が明確な評価差として現れにくく，評価結果の解釈が困難となる場合があるという課題が確認されている。

こうした問題意識は，ロボット倫理や科学技術社会論における議論とも接続する。呉羽は，責任主体になり得ない人工物に謝罪を行わせる行為の是非を問い，ロボットを通じて人間社会の謝罪慣行そのものを問い直す可能性を示している [4]. この観点は，謝罪を行う対話ロボットを性能評価の対象としてのみ捉えるのではなく，人間の評価行為や判断基準を可視化する装置として捉える視座を与える。

本研究では，謝罪を行う対話ロボットの性能向上を直接的な目的とするのではなく，これまでに実施してきた主観評価実験の経験をもとに，謝罪行動の評価がどのような点で困難となるのかを整理する。従来の評価手法で直面した問題点を明確化した上で，今後の実験的評価設計や評価尺度構成に向けた検討の前提を示すことを目的とする。

2　関連研究

謝罪を行う対話ロボットに関する研究では，失敗回復行動や対話戦略の一部として謝罪発話を組み込む試みが数多く報告されている。これらの研究では，謝罪表現の有無や文言の違いが，印象評価や信頼感に与える影響が検討されてきた [1, 2]. また，音声抑揚や表情変化を組み合わせたマルチモーダル表現が対話の自然さを高める可能性も示されている。

一方で，謝罪を社会的行為として捉えた場合，その評価は単純な好悪判断に還元できない。ロボット倫理の分野では，責任主体にならない人工物に謝罪を行わせることが，人間側の責任意識や規範理解に影響を及ぼす可能性が指摘されている [4]. これらの

議論は規範的観点から展開されることが多く，実際の評価場面において人間がどのように謝罪ロボットを評価しているかを実験的に整理した研究は限られている。

本研究は，謝罪対話ロボットそのものの振る舞いを評価するのではなく，評価行為そのものに着目し，主観評価がどのような点で成立しにくくなるのかを整理する点に特徴がある。

3　謝罪対話ロボットシステム

本研究で用いた対話主体は，課題提出の遅延などの過失を犯した学生を想定した三次元 CG アバターであり，ユーザに対して謝罪を行う対話ロボットとして設計した。物理ロボットではなく，社会的役割を明示的に背負わせた対話主体である点に特徴がある。

本アバターは，謝罪という社会的行為に着目し，言語表現，音声的特徴，表情変化を統合したマルチモーダルな振る舞いを生成する。本研究では，このロボットを評価対象そのものとするのではなく，評価を引き起こす装置として位置づける。

4　評価経験の整理

4.1　従来の評価方法

これまで我々は，謝罪対話後に質問紙を提示し，好感度，誠実さ，納得感などに関する主観評価尺度を用いた実験を実施してきた。条件間比較により，謝罪表現や非言語的振る舞いの違いが評価結果に与える影響を検討した。

しかし，実験結果では想定したような条件差が明確に現れない事例が多く確認された。

4.2　評価が明確にならなかった具体例

例えば，責任を明示する謝罪表現と簡潔な謝罪表現を比較した実験では，平均値にわずかな差は見られたものの，評価者間のばらつきが大きかった。また，表情や音声トーンを付与した条件と発話のみの条件を比較した際にも，どちらが望ましいかについて評価者の判断は分かれた。

一方で，自由記述には「丁寧で誠実に感じる」「形式的に謝っているだけに見える」といった相反する評価が多数記述され，数値化が困難な判断基準の存在が示唆された。

4.3　問題点の整理

これらの経験から，評価基準が評価者ごとに大きく異なる点，謝罪を一時点の発話として評価していた点，判断を強制する設計が謝罪行為の性質と必ずしも適合していない可能性が示唆された。従来の主観評価手法では，謝罪行動の差異や評価者の判断過程を十分に捉えきれない場合があると考えられる。

5　評価困難性を踏まえた研究の位置づけ

以上の結果を踏まえると，評価困難性は評価手法の工夫のみで解消される問題ではなく，謝罪という行為そのものがもつ性質に起因する可能性が高い。謝罪は，過失の受け止め方，相手の感情への配慮，関係修復への姿勢を含む社会的行為であり，その受け止め方は評価者の経験や価値観，文脈理解に強く依存する。このため，評価者間で判断基準が共有されにくく，一義的な優劣判断が困難となる状況が生じ得る。

本研究の目的は，新たな評価尺度や評価手法を直ちに提示することではない。今後は，従来の主観評価尺度が前提としてきた「評価基準の共有」や「一時点での判断」が成立しにくい状況を整理し，謝罪行動の評価がどのような点で行き詰まるのかを明らかにする必要があるだろう。

6　評価尺度構成に向けた取り組み

謝罪行動の誠実性をどのように測定するかという課題については，近年，評価尺度の構成を目的とした基礎的な検討が進められている。これらの取り組みでは，大規模な主観評価実験を直ちに実施するのではなく，「誠実な謝罪」と判断される際に，人がどのような観点に着目しているのかを明らかにすることが重視されている。

具体的には，対面での謝罪場面を想定し，誠実と感じる謝り方について自由に記述してもらう調査を通じて，評価尺度項目の候補となる判断観点の整理が行われている。このような調査では，誠実性をあらかじめ単一の概念として定義するのではなく，評価者自身の言語化を手がかりとして，多面的な判断基準を抽出することが意図されている。

本稿で整理した評価困難性の事例は，こうした尺度構成の前段階における問題意識と整合するものである。すなわち，謝罪行動を一義的に評価すること

の難しさを事例として示す本研究の知見は，今後，謝罪行動の誠実性を測定する評価尺度を構成する際に検討すべき前提条件を明らかにするものであると考えられる。

7 おわりに

本稿では，謝罪を行う対話ロボットに対する主観評価実験の経験をもとに，従来の評価尺度では解釈が困難となる事例を整理した。評価結果が数値として一貫して現れにくい点や，評価者ごとに判断基準が異なる点は，謝罪が社会的行為として多義的かつ文脈依存的であることに起因すると考えられる。

本研究は，謝罪対話ロボットの性能向上を直接的に目指すものではなく，評価が成立しにくい状況を事例として整理することで，今後の評価研究や評価尺度構成に向けた基礎的検討を支える位置づけにある。今後は，評価者が誠実性を判断する際に用いている観点を段階的に整理し，謝罪行動の特性を踏まえた評価尺度の構成について検討を進める予定である。

参考文献

[1] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, Gracefully Mitigating Breakdowns in Robotic Services, In *Proc. of HRI 2010*, pp. 203–210, 2010.

[2] D. Cameron, S. de Saille, E. C. Collins, et al., The Effect of Social-Cognitive Recovery Strategies on Likability, Capability, and Trust in Social Robots, *Computers in Human Behavior*, Vol. 114, 106561, 2021.

[3] 酒造正樹，小野田凌也，学生によるプロンプトチューニングを用いた謝罪するロボットのもたらす教育効果，言語処理学会第 31 回年次大会（NLP2025），P-35，2025.

[4] 呉羽真, ロボットで社会を揺さぶる〈クリティカル・ロボティクス〉，第 41 回日本ロボット学会学術講演会，2F3-01，2023.