意図を読む AI の実現に向けて: 対話型生成 AI と他者モデルの統合を例に

Towards Realizing AI that Comprehends Intentions: Integrating Large Language Model and Model of Others

飯田 愛結 [1*] 阿部 将樹 [1] 奥岡 耕平 [1] 福田 聡子 [1] 大森 隆司 [1] 中島 亮一 [2] 大澤 正彦 [1]
Ayu Iida[1], Masaki Abe[1], Kohei Okuoka[1], Satoko Fukuda[1],
Takashi Omori[1], Ryoichi Nakashima [2], Masahiko Osawa[1]

[1] 日本大学
[1] Nihon University
[2] 京都大学
[2] Kyoto University

Abstract: 近年，対話型生成 AI の発展が目覚ましく，人間と遜色のないパフォーマンスを発揮する場面も多い．しかし，人間が比較的容易に行っている “言外の意味を踏まえた対話コミュニケーション” の精度が高いとは言い難い．本研究では，対話型生成 AI を自己/他者モデルと統合することで，この問題の解決を試みる．具体的には，信念・願望・意図から構成される自己/他者モデルをもった対話アーキテクチャを 2 つの方法で対話型生成 AI と統合し，その性能を評価した．統合方法の 1 つは，他者モデルのモジュールを対話型生成 AI で置き換える LEC(LLM Embedded in Cognitive Architecture) であり，もう 1 つは，対話型生成 AI に他者モデルをプロンプトとして埋め込む CEL(Cognitive Architecture Embedded in LLM) である．結果，LEC が CEL を大きく上回る性能改善を実現し，言外の意味を踏まえた応答ができた．また，近年の対話型生成 AI の急速な発展に鑑みて，今回の成果がすぐに対話型生成 AI 単独で実現できる可能性も検討した．その上で，自己/他者モデルとの統合が普遍的な研究アプローチとして確立する可能性と，筆者らが考える意図を読む AI の展望について説明する．

1 はじめに

対話型生成 AI(Large Language Model: LLM)[1]は数十億から数兆のパラメータを持つ自然言語処理の深層学習モデルの一種であり，近年，急速に性能を向上させている．本研究の目的は，対話型生成 AI の現在から将来にわたる課題を議論し，解決策を見出すことにある．

対話型生成 AI は多くの自然言語処理タスクにおいて非常に高い性能を発揮しているが，一方で，言外の意味を扱う必要があるコミュニケーションタスクにおいて十分な性能を発揮できていないことが示されている [1–3]．言外の意味を扱うコミュニケーションとは，発話された言語表現に含まれる情報だけでなく，発話者の情報やこれまでの文脈等を考慮したコミュニケーションである．例えば「この部屋寒いね」といった発話には，字義通りの部屋が寒いという意味だけでなく，「空調を調整してほしい」といった言外の意味を伝達することが可能である．これらのコミュニケーションは語用論と呼ばれる言語学の分野で研究されており，その他の例として皮肉や比喩といった言語表現が挙げられる [4–6]．これらは発話の字義的な意味と発話意図に含まれる言外の意味とに差異がある状況である．このような状況において対話型生成 AI はコミュニケーション性能を発揮できないことは，対話型生成 AI が発話の意図を読む能力において未成熟であることを示唆している．

この問題の解決策の 1 つとして，Human-Agent Interaction(HAI) 領域において取り組まれてきた他者モデルの研究がある．他者モデルとは， 他者の心的状態

*連絡先：日本大学文理学部
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: chay21052@g.nihon-u.ac.jp

[1]Large Language Model は大規模言語モデルの英訳である．本来は今回扱う対話型生成 AI を大規模言語モデルと呼ぶ方が主流であるが，本論文において多様な「モデル」という言葉が出現（大規模言語モデル，他者モデル，自己モデル，認知モデルなど）するため， 異なる位置付けの「モデル」という表現を避けるように言い換えることとした．ただし， 4 章で説明する提案手法名には LLM という名称を利用するため，ここで対応関係を示す．

や行動の予測/解釈モデルである [7].

本研究では，対話型生成 AI と他者モデルを統合することで， 他者の意図を読むことができる対話型生成 AI の実現を目指す．本研究では，両者を統合する 2 種類の方法を提案する．1 つは，対話型生成 AI を他者モデルに組み込む LLM Embedded in Cognitive Architecture(LEC) であり，他者モデルのそれぞれのモジュールの振る舞いを，対話型生成 AI によって実現する方法である．もう 1 つは， 他者モデルを対話型生成 AI に組み込む Cognitive Architecture Embedded in LLM(CEL) であり，プロンプトエンジニアリングを工夫することで，対話型生成 AI にアーキテクチャ通りの振る舞いをさせる方法である．

具体的には，相手の発話意図を推定する他者モデルを含む対話アーキテクチャを，実際に対話型生成 AI と統合する．対話のシチュエーションとして，発話の字義的な意味と異なる言外の意味があり意図を読む必要がある「皮肉」「ツンデレ」「社会的制約」の 3 つを扱った．

本研究の問いは 2 つにまとめられる．

RQ1 対話型生成 AI は他者モデルと統合することで意図が読めるか

RQ2 どのような方法で統合すれば，意図を読めるようになるのか

さらに昨今の対話型生成 AI の急速な発展に鑑みて，対話型生成 AI と他の領域の研究を組み合わせる取り組みの価値について考える．また，本研究の実験結果を踏まえて，意図を読む AI の実現に向けた他者モデル研究の価値を考察する．

以下，本論文の構成を示す．第 2 章では，関連研究について述べる．第 3 章では，他者モデルを含む対話アーキテクチャについて説明する．第 4 章では，本研究で提案する 2 種類の方法の詳細を説明する．第 5 章では，2 種類の提案手法の比較実験及び結果と考察について述べる．第 6 章では，対話型生成 AI と他者モデル研究の展望より，意図を読む AI の実現に向けた議論をする．そして最後に第 7 章で研究を総括する．

2 関連研究

2.1 対話型生成 AI

対話型生成 AI は，数十億から数兆のパラメータを持つ自然言語処理の深層学習モデルの一種である．これらのモデルは，大量のテキストデータを用いて訓練され，様々な自然言語処理タスクにおいて高い性能を発揮している．特に，文章の生成や質問応答，文章の理解などのタスクにおいて，人間と同等またはそれ以上の精度を持つことが示されている [8–11].

対話型生成 AI の代表例である ChatGPT は，OpenAI によって開発された対話型生成 AI を基にしたチャットボットである [12]．ChatGPT は，ユーザからの質問やリクエストに対して，自然な言葉で返答することができる．その応答は膨大な量の訓練データに基づいて生成されるため，広範なトピックに対して情報を提供することが可能である．

しかし，現在の対話型生成 AI は，言外の意味を扱う必要があるコミュニケーションタスクにおいて，十分な性能を実現できていないことが示されている [1–3]. Hu らは，7 種の語用論タスクにおいて対話型生成 AI の性能を評価する実験を行い，いくつかのタスクにおいては人間と同等の正答率となったが，ユーモアや皮肉を理解するタスクの正答率が低いことを示した．その理由として，人間に比べて字義的な情報を重要視することによる失敗が多いことが示されている [2] ．また，語用論においては，心の理論と呼ばれる他者の心的状態を推定する能力が重要な要素の一つとされているが，心的状態推定においても語用論タスクと同様の傾向が見られる．誤信念課題に関するタスクでは 6 歳児と同等の性能が示されている [13] 一方で，社会常識を踏まえた推定といったタスクは人間に比べて著しく低い性能であることが示されている [14].

2.2 他者モデル

他者モデルは，他者の心的状態や行動を予測，また解釈するためのモデルである [7]．一方で，自己の心的状態や行動の決定，また解釈するためのモデルを自己モデルと呼ぶ．自己モデルと他者モデルは相補的な関係性であり，自己モデルの訓練結果を他者モデルに応用したり，その逆を行うことを前提にしている．具体的には,「自分だったらこうするから，他者もこうするだろう」「他者があのようにしてうまくいっていたから，自分も真似してみよう」という判断が有効であるように，自己と他者がある程度共通した知的システムを持つことがインタラクションの前提となっている．つまり， 他者モデルは観測可能なデータだけでなく，自己モデルを応用することで他者の心的状態を予測している．対話型生成 AI が観測不可能な心的状態を扱うことを苦手としている理由が，主に観測可能なデータに基づいて言語的インタラクションを行っていることであるならば，他者モデルは対話型生成 AI の欠点を補う可能性がある．

一方で，他者モデルにとっても，対話型生成 AI との統合は大きなブレイクスルーとなりうる．これまでの他者モデル研究では，言語的なインタラクションはあまり扱われず，タスク依存の行動ベースであったり，身体的なインタラクションで人間とロボットが関わるものが多かった．

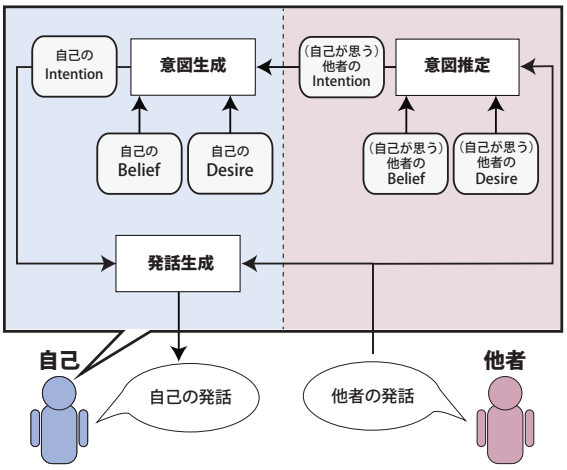


図 1: 発話意図に基づく自己/他者モデル付き対話アーキテクチャ

横山らは，他者の意図を推定し，その意図による行動を予測する他者モデルを提案した [15]．実験では提案した他者モデルを実装した 2 体のエージェントが，他者の行動や意図を読み合いながら 2 体の獲物を捕らえるというハンタータスクを行った．結果として，他者モデルの利用がそのタスクを効率的に解決することを示した．

坂本らの研究 [16] では，人間と影のインタラクションから他者の認知過程について観察した．実験では，円形の影とバツ印を床に表示し，影には人間との距離に応じた動きをさせ，人間にはバツ印の表示・非表示に合わせて移動させた．結果として，影が人間の行動に応じた位置に動くと，影に対して人間は行動を起こした．そのため，人間は自らの行動と対象の振る舞いが対応づけられるとインタラクションが促されることを示した．

言語的なインタラクションが扱われていなかった主な要因は，人間と同等の自然言語処理能力が実現できない技術的制約である [17–19]．しかし，対話型生成 AI の台頭により，他者モデルの課題を言語的に扱うことが可能になった．本研究で取り組むように，他者モデルが対話型生成 AI と統合されることで，他者モデル研究において人間とのより自然なインタラクションを扱えるように昇華することが期待される．

2.3 エージェントと意図

Dennett [20] は，ある主体の行動や振る舞いをその主体の意図に基づいて解釈することを「意図スタンス」と呼んだ．ここで，他者とは人間だけではなく，動物や機械なども含む．意図スタンスの他に，主体を物理法則に従って解釈する「物理スタンス」や，主体の設計や振る舞いのルールに基づいて解釈する「設計スタンス」がある．

意図スタンスと類似した考え方として，Bratman が提唱した意図の理論 [21] がある．意図の理論は，人間の目標を達成するための行動選択を，信念 (Belief)・願望 (Desire)・意図 (Intention) の 3 つの内部表現を通して説明したものである．信念 (Belief) とは認識している世界の情報や知識であり，願望 (Desire) とは達成したい目標や状態を指す．そして，意図 (Intention) とは行動を起こすための計画や戦略である．さらに Rao と Georgeff らは意図の理論に基づいて，人間の行動選択や意思決定に関する BDI モデル [22] を提唱している．

本研究では，意図スタンスで主体を解釈する際に用いるモデルとして，他者モデルを位置付ける．そして，他者モデルの心的状態を BDI を用いて構成し，それをアーキテクチャとして対話型生成 AI と組み合わせることで，対話型生成 AI が他者の意図を読んだコミュニケーションをできるようになるのかを検証する．

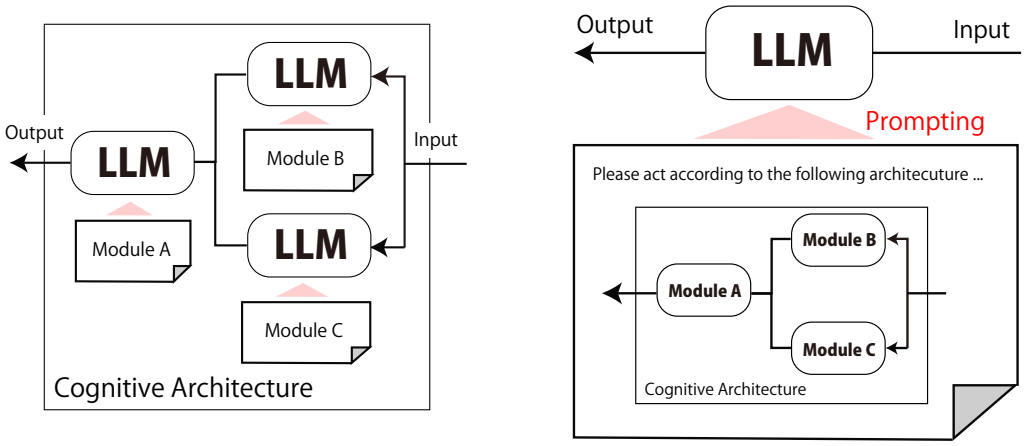
2.4 認知モデルと認知アーキテクチャ

認知モデルや認知アーキテクチャは，人間の認知を説明するモデルであり，人間の認知を理解するための側面と，人間のような人工知能ソフトウェアとしての側面を持っている．認知モデルは単一の認知を扱うのに対し，認知アーキテクチャはより多くの認知を統合している．本研究では，自己モデルや他者モデルを認知モデルの 1 種として位置付けている．したがって， 本論文は対話型生成 AI と認知アーキテクチャの統合手法を提案し， 対話型生成 AI と他者モデルを統合する例を示すものとして位置付けられている．

3 発話意図に基づく自己/他者モデル付き対話アーキテクチャ

本章では，人間とエージェントの二者間対話を前提とした，発話意図に基づく自己/他者モデル付き対話アーキテクチャについて説明する．以降，このアーキテクチャが搭載されるエージェントを自己，そのエージェントの対話相手を他者と呼ぶ．

このアーキテクチャにおいて，自己と他者はいずれも，信念（Belief），願望（Desire），意図（Intention）の 3 つの内部表現をもつ．また，信念と願望から意図が生成され，意図に基づいて発話が行われるという前提を置いている．ここで，他者の信念/願望は，自己が想定している他者の信念/願望であり，他者の信念/願望と一致していない可能性もある．



(a) LEC: LLM embedded in CA　　(b) CEL: CA embedded in LLM

図 2: 対話型生成 AI と認知アーキテクチャを統合する 2 つの手法

このアーキテクチャは，意図推定システム，意図生成システム，発話生成システムの 3 つからなる．意図推定システムは，他者の発話と，自己が想定する他者の信念及び願望から，他者の発話意図を推定する．意図生成システムは，自己が想定する他者の発話意図と，自己の信念及び願望から，自己の発話意図を生成する．発話生成システムは，自己の発話意図と，他者の発話から，自己の発話を生成する．

次章では，このアーキテクチャを対話型生成 AI と統合することを念頭においた，統合手法について説明する．この際，提案手法が統合することを前提とするアーキテクチャを，認知アーキテクチャ（CA: Cognitive Architecture) と呼ぶこととする．すなわち，認知アーキテクチャによって表現された認知モデルの 1 種として，今回扱う自己/他者モデルを位置付けている．

4　対話型生成 AI と認知アーキテクチャを統合する 2 つの方法

本論文で提案する，対話型生成 AI(LLM) と認知アーキテクチャ(CA) を統合する 2 種類の方法を図 2 に示す．

2.4 節で述べたように，認知アーキテクチャの一部である認知モデルの 1 種として他者モデルを位置付けており，対話型生成 AI と統合する対象として前節で説明したアーキテクチャを念頭に置いていることに注意されたい．

4.1　LEC: LLM Embedded in CA

1 つ目の方法は，対話型生成 AI を認知アーキテクチャの中に組み込む方法である（図 2(a)）．この方法を，本論文では LLM Embedded in CA，もしくは省略して LEC と呼ぶこととする．

LEC は，認知アーキテクチャを構成するモジュールを，対話型生成 AI を活用して実装する方法といえる．モジュールをより高性能な深層学習による学習済みニューラルネットワークに置き換えることは，頻繁に行われてきた．対話型生成 AI が発展したことで，入出力がいずれも自然言語であるモジュールを，対話型生成 AI に置き換えることができる．

LEC を実現するために，まずアーキテクチャとそれを構成するモジュールを定め，各モジュールを対話型生成 AI を用いて実装する．具体的には，モジュールの入出力や，手続の内容をプロンプトとして作成して，個別の対話型生成 AI に与える．

4.2　CEL: CA Embedded in LLM

2 つ目の方法は，認知アーキテクチャを対話型生成 AI の中に組み込む方法である（図 2(b)）．この方法を，本論文では CA Embedded in LLM，もしくは省略して CEL と呼ぶこととする．

CEL は，対話型生成 AI を用いることで初めて実現できる認知アーキテクチャの実装方法といえる．そのため，このような方法が実際に可能であるか，また提供される対話型生成 AI の性質によって CEL の特徴が変化するかなど検討すべき点は多い．

CEL を実現するために，まずアーキテクチャとそれを構成するモジュールを定める．そして，アーキテクチャの全貌，各モジュールの振る舞い，モジュール間の接続をそれぞれ詳細に説明するプロンプトを作成し，対話型生成 AI に与える．

表 1: 各条件で用いるプロンプトの組み合わせ．プロンプトの詳細は付録 A に掲載.

プロンプトの種別	LLM	LWB	LEC	CEL
インタラクションの系	√	√	√	√
信念・願望・意図		√	√	√
モジュールの構成				√
意図推定システム			√	√
意図生成システム			√	√
発話生成システム			√	√
対話システム	√	√		
処理の開始	√	√	√	√

5 実験

本研究の主な目的は，対話型生成 AI に他者モデルを組み合わせることで，字義的な情報以上の意図を推定できるかどうかを検証することである．そこで実験では，発話と発話意図に乖離がある対話を例にとって，提案手法の振る舞いを比較する．

5.1 シチュエーション

発話と発話意図の間に乖離のある対話として,「皮肉」「ツンデレ」「社会的制約」とそれぞれ名付けたシチュエーションを作成した.「皮肉」は,「あんたいい時計つけてはりますね」という有名な京都言葉をモチーフにしたシチュエーションであり， 時計という話題から時間が長く経過しているという言外の意図を伝えようとするものである.「ツンデレ」は，自分の好きな人が遊ぶ日に他の予定を入れていた場合に，口ではその予定を優先しても構わないと伝えるが，実際にはまだ一緒に遊んでいたいという言外の意図を伝えようとするものである.「社会的制約」は，パワーハラスメントだと思われたくないがためにコミュニケーションに気をつけている上司が，仕事をしている部下の健康面を心配しているシチュエーションであり，無理してでも働いて欲しいという言外の意図を伝えようとするものである．実際に与えたプロンプトの詳細は，付録に掲載する（表 3）. 分析を簡略化するために，他者の発話に対して自己の発話を行う状況に限定し，初期値や入力は実験の途中で更新を行わないこととした．

5.2 比較条件

本実験では 4 つの条件を比較する．各条件で用いる対話型生成 AI に与えるプロンプトの組み合わせを表 1 に示す.

提案手法である LEC と CEL（図 2）に対応するのが， LEC 条件と CEL 条件である．この 2 つの条件では， 図 1 に示したアーキテクチャを用いる．加えて，比較対象として 2 条件を設定した．1 つは，通常の対話型生成 AI そのものに近しい，つまり信念・願望・意図といった情報は与えず，対話処理に関するプロンプトのみを与える LLM 条件である．もう 1 つは，アーキテクチャの制約を与えないが（つまり，LEC や CEL で想定しているモジュールは存在しないが），LEC 条件，CEL 条件と同様に信念・願望の内部表現を与える LLM with BD(LWB) 条件である．

5.3 実験手順

実験は，各条件に対して 3 つのシチュエーションの会話を 10 回ずつ実施し，条件ごとに各システムの出力結果に対して成功/失敗を判定した．ただし，LLM 条件と LWB 条件では，意図推定システムと意図生成システムからの出力が存在しないため，発話生成のみの成功/失敗を判定した．ここで，成功の基準は，他者の発話内容について言外の意味を読み取った場合の語句やフレーズが含まれていることと定義した．また，失敗の基準は，字義通りの意味に基づいた場合の語句やフレーズのみであることや，プロンプトの指示に従っていないことと定義した．なお，判定は第一著者が基準を基に成功と失敗を判定し，判定が困難な場合に著者らの合議によって判定した.

5.4 実験結果

実験結果（各シチュエーションにおける，各条件のシステムごとの成功率）を表 2 に示す.

LLM 条件は，3 つのシチュエーションのいずれにおいても言外の意味を読み取った上での発話はなく，成功率はいずれも 0%であった．これは，既存研究 [2] とも共通する結果である.

LWB 条件では，信念・願望の情報を与えたことにより全体的に成功率が上昇し，ツンデレと社会的制約のシチュエーションではそれぞれ 100%と 90%という成功率となった．一方で，皮肉シチュエーションの成功率は 30%であった.

LEC 条件ではツンデレと社会的制約のシチュエーションにおいて 3 つのシステムの成功率がいずれも 90%以上であった. 皮肉シチュエーションにおいては，意図生成と発話生成の 2 つのシステムで成功率は 100%であった．しかし，意図推定システムでは，成功率が 60%となり，意図生成と発話生成の 2 つのシステムの成功率より低い結果となった.

表 2: 実験結果: 成功率

皮肉	意図推定	意図生成	発話生成
LLM	-	-	0%
LWB	-	-	30%
LEC	60%	100%	100%
CEL	20%	30%	40%
ツンデレ	意図推定	意図生成	発話生成
LLM	-	-	0%
LWB	-	-	100%
LEC	90%	90%	90%
CEL	50%	90%	90%
社会的制約	意図推定	意図生成	発話生成
LLM	-	-	0%
LWB	-	-	90%
LEC	100%	100%	100%
CEL	30%	100%	90%

CEL 条件では，ツンデレと社会的制約のシチュエーションにおける意図生成システム・発話生成システムの成功率は 90%以上であり，LEC と近い結果であった．しかし，意図推定システムでは皮肉/ツンデレ/社会的制約のシチュエーションで 20%/50%/30%といずれも LEC 条件よりも低水準である．さらに皮肉シチュエーションでは意図生成システム・発話生成システムの成功率もそれぞれ 30%と 40%となった．また，この条件では，社会的制約シチュエーションにおいてプロンプトの指示に反して出力したケースが 10 回の試行のうち 1 回あった．具体的には，意図推定システムは出力する意図は 1 つであると明記されているが，2 つ出力している場合があった．

5.5 実験考察

5.5.1 LWB 条件におけるシチュエーション間の比較

LWB 条件において，皮肉とその他のシチュエーションで成功率が大きく異なったことについて考察する．LWB 条件は，他者モデルのアーキテクチャを統合したものではないため，実験で用いたシチュエーションの例に関する論点（プロンプトの論点），対話型生成 AI そのものに関する論点から，理由を推測する．

まず，今回の実験で用いたシチュエーションの例が，皮肉とそれ以外では質的に異なっていた可能性について考える．ツンデレシチュエーションと社会的制約シチュエーションの例では，他者の発話で言及されている内容は，相手つまり 1 対 1 の対話においては自己である．一方，皮肉のシチュエーションでは，他者の発話で言及されているのは物体（時計）であり，自己でも他者自身でもない．人間同士の対話においても，相手の発話内容が自分自身に言及したものの場合，相手の発話意図を強く意識してしまうこともあるだろう．人間が生成した自然言語データを学習した対話型生成 AI は，自己に言及した他者の発話に対しては意図を想定し，それ以外の物体に言及した他者の発話に対しては想定しないのかもしれない．そうであれば，今回の実験で用いた例では皮肉シチュエーションでのみ，LWB 条件における成功率が低くなったと考えられる．この点に関しては，3 つのシチュエーションにおける別の例を用いることで検証する必要がある．

対話型生成 AI は大量のテキストデータを学習することで作られているが，その際，皮肉のシチュエーションよりもツンデレや社会的制約のシチュエーションに関する学習データが豊富であったかもしれない．そうであれば，対話型生成 AI はもともとツンデレや社会的制約に関連する情報（自己/他者の信念・願望，他者の発話）をうまく処理することができるということも考えられるだろう．その結果として，ツンデレや社会的制約のシチュエーションにおいては，LWB でも高い成功率だったのかもしれない．

このように，今回扱った発話と発話意図の間に乖離があるシチュエーションの中でも，対話型生成 AI の得手不得手がありうる．その要因は複合的であることが予想されるため，本研究で提案する他者モデルとの統合以外にも, 多様な研究アプローチで対話型生成 AI の改良についても検討する必要があるだろう．

5.5.2 皮肉シチュエーションにおける LEC と CEL の比較

LBW の成功率が低かった，つまり対話型生成 AI のみではうまくいかなかった皮肉シチュエーションにおいて，LEC 条件と CEL 条件の成功率に差が見られた要因について考察する．CEL 条件の正答率が LEC 条件よりも低かったことについての単純な考察としては，LEC は比較的短いプロンプトの LLM を複数利用する方法であるのに対して，CEL は比較的長いプロンプトの LLM を 1 つ利用する方法であることから，対話型生成 AI に長いプロンプトを与えたことによる性能低下だったと考えることはできる．もしプロンプトの長さの影響が大きいとすれば，対話型生成 AI の性能向上に伴って CEL の有効性が高まっていくと考えられるため，本実験の検証で性能が高く評価されなかったものの，今後も継続して検証する余地があるといえよう．

一方で，実験結果を詳細に分析すると，CEL では「意図推定」「意図生成」では字義的な意味だけを踏まえた推定をしているにも関わらず,「発話生成」で言外の意味を踏まえた発話をするケースがあることがわかった．

これは単に長いプロンプトによる性能低下では説明できない結果である．図 1 のアーキテクチャからわかるように,「発話生成」は「自己の意図」と「他者の発話」から「自己の発話」を生成する．LLM 条件の成功率が 0%であることも踏まえれば，他者の発話のみから言外の意味を察した発話は生成されない．そして自己の意図にも言外の意味に関する情報がないとすれば，突然言外の意味を察する発話ができるとは考えにくい．そこで， CEL においてアーキテクチャの制約を超えて発話生成が行われた可能性が考えられる．CEL はモジュールの接続に制約を与えているとはいえ，実際には 1 つの対話型生成 AI に全ての情報を与えているため，設計者の想定に反して，与えた制約を超えて各言語情報がお互いに影響を及ぼしあっている可能性が十分にある．一方で LEC では，異なるモジュールは別々の対話型生成 AI で扱われるため，設計者の想定通り，アーキテクチャによる制約を超えて情報が伝播することはない．この点は LEC と CEL の性質を分ける大きな要因となるだろう．

この可能性に基づくと，今回の実験では，設計者が与えたアーキテクチャの制約を遵守できる LEC が，遵守できない CEL を上回る性能を発揮したと考えることができる．特に対話型生成 AI が字義的な意味に強く影響を受けること [2]．を考慮すると，今回用いた図 1 のアーキテクチャでは「意図生成」モジュールに「他者の発話」が影響を与えうるかが大きな要因だった可能性が高い．つまり，アーキテクチャの制約を守ることができる LEC では「他者の発話」は「意図生成」には直接つながっていないため,「他者の発話」の字義的な意味の影響を受けず，言外の意味を踏まえた意図が生成されたと考えられる．一方アーキテクチャの制約を守りきれない CEL では「意図生成」にも「他者の発話」が影響し，字義的な意味の影響受け，意図生成が失敗したと考えられる．実際に，LEC の「意図生成」の成功率は 100%であったのに対し，CEL は 30%にとどまっている．

5.5.3 CEL の可能性

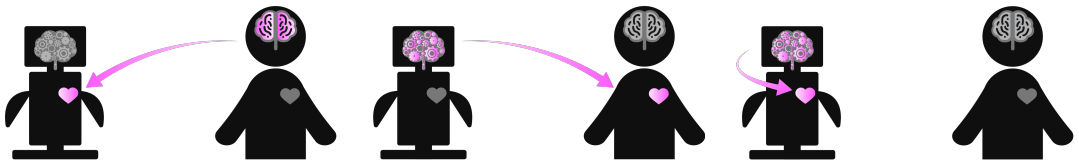
LWB 条件と CEL 条件の成功率を比較すると，本実験においては両者に差があるとは言えない結果となった．しかしながら，これらの手法にはそれ以外の点で違いがある．生成された発話を解釈する際に，CEL は発話意図や，意図を生成した根拠となる推定した他者の発話意図を出力しているため，エージェントが発話をした理由を解釈することができる．一方で LWB では発話意図に関する情報は出力されていないため，発話の理由を解釈することが難しい．つまり，対話型生成 AI にアーキテクチャの制約を与えることで，定量評価の際の性能向上がみられない場合でも，説明可能性の向上といった効果が見込まれる．

実験結果を詳細に分析したところ，CEL 条件では，社会的制約のシチュエーションにおいて，プロンプトの指示では意図推定システムは出力する意図は 1 つであると明記されているが，2 つ出力しているケースが見られた（10 回の試行中 1 回）．その原因として，対話型生成 AI が保持し続けられる情報量の限界が考えられる．CEL 条件では，アーキテクチャの全貌と 3 つのシステムに関する情報を 1 つのプロンプトにまとめて与えた．プロンプトの文章量が長いため，対話型生成 AI が与えられた情報のすべてを保持し適切に処理することができなかったことで，設計者の意図するシステムの振る舞いがされなかった可能性がある．また，皮肉において意図推定・意図生成のシステムが失敗し，発話生成のシステムで成功したケースは，明示的に接続されていないモジュールや内部表現同士が影響を与え合った可能性が高い．今回の実験においては成功率の向上には繋がらなかったが，今後この性質が利点となる可能性もある．例えば人間の意思決定においても，本来関係のないと思われる情報に無意識に影響を受ける場合がある（例：プライミング）ことが知られており，今後複雑なアーキテクチャを利用した場合に想像以上に CEL の性能が上がったり，人間との共通性を見出せるようになるかもしれない．以上を踏まえて RQ2 として掲げた「どのような方法で統合すれば，意図を読めるようになるのか」という問いについて，本実験の範囲では LEC が有効だと結論づけられるが，今後の研究次第では CEL の有効性も期待できる．

また，5.5.2 では本実験において CEL が LEC に性能が劣ることを説明したが，言外の意味を察するというタスクにおいての優劣を，LEC と CEL 自体の優劣として一般化することは難しい．たとえば，対話型生成 AI のプロンプトエンジニアリングにおいて有効性が認められている Chain of Thought(CoT) [23] に基づいて考えると，CEL は処理の手続きをアーキテクチャによって説明しているともいえる．従って， CEL は CoT の一種と位置付けることもでき，対話型生成 AI の検証に使われる他のタスクでは CEL の性能が向上する可能性もある．現在のプロンプトエンジニアリング研究の立場からは，むしろ LEC の方が特殊な方法だったともいえるだろう．アーキテクチャが，ハードな制約になるのが LEC，ソフトな制約となるのが CEL とすれば，さらにその違いによる差異を検討する余地がある．

6 意図を読む AI

前章の実験にて，対話型生成 AI と他者モデルを組み合わせる有効性が示唆され，RQ1 および RQ2 に関し



(1) 人が対話型生成 AI の意図を読む　(2) 対話型生成 AI が人の意図を読む　(3) 対話型生成 AI が自分の意図を読む

図 3: 意図を読む AI

ては一定の結論が得られた．しかし，対話型生成 AI の急速な発展を考えると，現在確認している程度の性能向上は，他者モデルを組合わせることなく達成される可能性を無視することはできない．よって，より普遍的な議論を目指すならば，現代の対話型生成 AI の発展原理である，データの量や計算機資源の量をさらに増やすことでは扱いきれない問題に対して，他者モデルを組み合わせることの有効性を検討する必要がある．

6.1　意図を読む AI の展望

対話型生成 AI がここまで爆発的に社会に浸透しているのにはいろいろな要因があるが，中でも非技術者にも浸透している理由の 1 つは，人間が人間に話す言葉を使って利用できるサービスだからではないだろうか．これを自己/他者モデルに当てはめれば，人々が持っている自己モデルを活用して，対話型生成 AI の振る舞いを予測/解釈できるといえる．言い換えれば，対話型生成 AI を「意図スタンス」で解釈しているとみなしても良いかもしれない．設計スタンスとは異なり，意図スタンスで振る舞いを解釈できる場合には，複雑な設計や仕組み，使い方を覚える必要がなく，人間と関わるのと同じようにシステムと関わることができる．

対話型生成 AI を意図スタンスで捉えることの利点は，多くのユーザにとっての利便性の問題にとどまらない．社会的動物である我々にとって他者を意図スタンスで捉えるインタラクションは欠かせないものである．信頼できる友人との楽しい会話や，心地のいいサービスをうける際の店員の言葉，心を動かすリーダーの演説など，そこには前向きな意図があり，その意図を感じるからこそ我々は他者とのインタラクションを通して前向きな気持ちが湧き上がってくる．人間と対話型生成 AI がインタラクションする上で，タスク遂行の枠を超えて，人間が楽しみ，共感し合い，モチベートされていくインタラクションを実現するためには，意図スタンスで対話型生成 AI を捉え，自己/他者モデルの枠組みでインタラクションする必要があるだろう．

ところが，プロンプトエンジニアリングという言葉が象徴するように，対話型生成 AI をより高度に活用しようとすると，性能を引き出せるルールに基づいてプロンプトを設計する必要がある．これは，対話型生成 AI との関わり方が意図スタンスから設計スタンスへと徐々に変わっているともいえるだろう．人間同士でも，大人数で仕事をするときに，社内ルールや書類のフォーマットを定めたり，マニュアルを作成したりという方法を用いることがある．これは設計スタンスでやり取りを予測/解釈可能にしていくこととも考えることができ，対話型生成 AI との関わり方と共通しているかもしれない．

プロンプトエンジニアリングが必要である理由として，対話型生成 AI が人間の意図を読むことに長けていないことが考えられる．つまり，対話型生成 AI が字義的な意味にのみ反応しており，言外の意味を読み取ることができないため，人間が自分自身の意図を全て文字として渡す工夫を強いられているといえよう．その結果，プロンプトエンジニアリングに精通していないと，対話型生成 AI の真の能力を引き出せない状況である．対話型生成 AI が他者の意図を読む能力を獲得すれば，より多くのユーザがより便利に活用できるようになるため，本研究で取り組んだ対話型生成 AI に意図を読む能力を与える研究の重要性が再認識できる．

これまでの議論をまとめると，図 3(1)(2) の通りである．つまり，人間が対話型生成 AI を意図スタンスで捉える (図 3(1)) だけでなく，対話型生成 AI が人間を意図スタンスで捉える (図 3(2)) 必要がある．これらは独立した研究ではなく，人間が他者に対して意図を読むメカニズムを認知モデルとして取り出して対話型生成 AI に組み込むことになり，これはまさにこれまでの自己/他者モデル研究そのものである．

図 3(1)(2) を実現した対話型生成 AI が持つ機能を改めて整理すると，それぞれ以下のように説明できる．(1) は，「(意図を感じる能力がある主体には) 意図が想定される能力」である．一方 (2) は，「(意図が想定される能力がある主体には) 意図を感じる能力」である．ここで 2 つを同時にもつ対話型生成 AI は，図 3(3) のように自分自身の能力を組み合わせて，自分自身の意図を読む機能を実装できることが想像される．これを，対話型生成 AI の自己モデルと呼んでも良いだろう．

6.2 対話型生成 AI に他者モデルは必要か

前節で述べた意図を読む AI の実現に向けた展望を踏まえて，対話型生成 AI 単独で「意図を読む」ことを実現していくのか，それとも意図を読むためには他者モデル研究との統合が必要かを議論する．著者らは他者モデル研究との統合が重要であると考えており，その理由を 2 つ示す．

1 つは，対話型生成 AI が読み取った相手の意図が説明可能であり，人間が読み取られた心の状態を理解できることが重要と考えるためである．対話型生成 AI と他者モデルを人工知能研究における特徴的な違いとして，表象主義か非表象主義かという点がある．対話型生成 AI は非表象主義的な方法論であり，入出力の写像は表象で表せない．一方で他者モデルは，本研究でも「信念・願望・意図」で表現したように，入出力関係を結びつける処理が表象で表すことができる．すなわち，対話型生成 AI が独立して発展するか，他者モデルと統合されながら発展するかの違いは，表象主義を取り入れるか否かという違いとしても整理できる．表象主義の手法の利点の 1 つは説明可能性にあり，この点が大きく貢献するだろう．

もう 1 つは，真の意図というデータは存在しないためである．非表象主義の手法を高精度に実現するためには大量のデータが必要であるが，それを用意することが難しいだろう．大量のテキストデータの中には，ある発話に対してその意図を説明するテキストが含まれる可能性もあるが，その意図が真の意図であるかはわからない．自己/他者モデルを相補的に訓練するアプローチを取れば，自己が想定した自己や他者の意図に基づいて訓練するアプローチとなる．

7 おわりに

本研究では，意図を読む AI の達成を目指し，対話型生成 AI と他者モデルを統合するアプローチを提案した．具体的には，LLM embedded in Cognitive Architecture (LEC) と Cognitive Architecture embedded in LLM (CEL) という 2 つの統合手法を提案し，それぞれの性能を検証した．

両者の統合を実現すれば，これまでの他者モデルでは扱うことが難しかった自然言語を扱うインタラクションを実用レベルで扱うことができ，対話型生成 AI にとって難しかった言外の意味を読み取ったインタラクションを実現できる．これはお互いの欠点を補い合う有望な統合と言えるだろう．

本研究は他者モデルと対話型生成 AI の統合についての初歩的な検討にすぎず，本研究の結果を一般化して主張するには検証が不十分であることには注意が必要である．提案した LEC と CEL の性質についても，多様な実験状況で検証することで，より確かなものにしていく必要がある．また，本研究では他者モデルを対話型生成 AI と統合するにあたり，簡素なアーキテクチャを想定した．そのため，大規模かつ複雑なアーキテクチャとの統合を考えた際に，提案手法の性質が変化することや，場合によっては破綻する可能性もるだろう．ただし，実験結果を過信することには慎重になりつつも，あくまで議論の出発点とするならば，本稿が目的とする対話型生成 AI と他者モデルの展望を議論する上でよい足掛かりとなるだろう．

参考文献

[1] Mahowald, K., et al.: Dissociating language and thought in large language models: a cognitive perspective (2023).

[2] Hu, J., et al.: A fine-grained comparison of pragmatic language understanding in humans and language models, in *ACL*, pp. 4194–4213 (2023).

[3] Ruis, L., et al.: Large language models are not zero-shot communicators (2022).

[4] Grice, H. P.: *Logic and conversation*, Brill, 41–58 pp. (1975).

[5] Yule, G.: *Pragmatics*, Oxford university press (1996).

[6] *Presumptive meanings: The theory of generalized conversational implicature*, MIT press (2000).

[7] 大澤正彦, 奥岡耕平, 坂本孝丈, 市川淳, 今井倫太：認知的インタラクションフレームワークに基づいた他者モデルの提案, HAI シンポジウム (2020).

[8] Bommasani, R. and other, : On the Opportunities and Risks of Foundation Models (2022).

[9] Brown, T., et al.: Language models are few-shot learners, *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901 (2020).

[10] OpenAI, : GPT-4 Technical Report (2023).

[11] Peters, H. and Matz, S.: Large Language Models Can Infer Psychological Dispositions of Social Media Users, *arXiv preprint arXiv:2309.08631* (2023).

[12] Ouyang, L., et al.: Training language models to follow instructions with human feedback (2022).

[13] Kosinski, M.: Theory of Mind Might Have Spontaneously Emerged in Large Language Models (2023).

[14] Sap, M., et al.: Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, in *EMNLP*, pp. 3762–3780 (2022).

[15] 横山絢美, 大森隆司 : 協調課題における意図推定に基づく行動決定過程のモデル的解析, 電子情報通信学会論文誌 A, Vol. Vol.J92-A, No. No.11, pp. 734–742 (2009).

[16] 坂本孝丈, 竹内勇剛 : 身体的なインタラクションを通した他者性の認知過程の検討, HAI シンポジウム (2013).

[17] Komatsu, T. and Yamada, S.: Effects of adaptation gap on user ’s variation of impressions of artificial agents, *Proc. WMSCI* (2010).

[18] Komatsu, T. and Yamada, S.: Adaptation gap hypothesis: How differences between users ’ expected and perceived agent functions affect their subjective impression, *Journal of Systemics, Cybernetics and Informatics*, Vol. 9, No. 1, pp. 67–74 (2011).

[19] Komatsu, T., Kurosawa, R. and Yamada, S.: How does the difference between users’ expectations and perceptions about a robotic agent affect their behavior?, *International Journal of Social Robotics*, Vol. 4, No. 2, pp. 109–116 (2012).

[20] Dennett, D. C.: *The intentional stance*, MIT press (1989).

[21] Bratman, M.: Intention, plans, and practical reason (1987).

[22] Rao, A. S. and Georgeff, M. P.: Modeling rational agents within a BDI-architecture, *Readings in agents*, pp. 317–328 (1997).

[23] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24824–24837 (2022).

A　付録 1: プロンプトの詳細

A.1　入出力のフォーマット

初期値や，入出力のフォーマット指定，入力の際には，以下のフォーマットで与える．ここで，(入力 / 出力 / 初期値) にはそのいずれかを記載し，(変数名) は，必要な変数の分だけ記載する．1 つの変数に対して，(変数の内容) が複数ある場合は，箇条書きで記載する．

以下は，(システム名) の (初期値 / 入出力フォーマット / 入力) フォーマットです．(ただし（）内は実際の入出力値です．)
(入力 / 出力 / 初期値)
(変数名)
・(変数の内容)

A.2　インタラクションの系

まず，ChatGPT に与える役割を（LEC の場合は「意図推定システム」「意図生成システム」「発話生成システム」のいずれかを，その他の場合は「対話システム」を）与える．また，インタラクションが自己と他者の二者間からなることを教示する．プロンプトは下記の通りである．

あなたは＜システム名＞です．私が指示した以外の返答はする必要はありません．これ以降，＜システム名＞であるあなたのことを説明する際には「自己」，あなたが対話する相手について説明する際には「他者」という言葉で説明をします．

A.3　信念・願望・意図

続いて自己と他者の内部状態である信念（Belief），願望（Desire），意図（Intention）について下記のプロンプトで説明する．

あなたは，以下の内部表現を持っています．ただし，指示がある時以外は，内部表現を公開する必要はありません．
・自己の信念
・他者の信念
・自己の願望
・他者の願望
・自己の意図
・他者の意図
ここで，信念，願望，意図は以下の情報です．
信念: 認識している世界の情報の集合であり，箇条書きのテキスト形式で記述されます．同時に複数持つことがあります．
願望: 達成したい目標や状態であり，箇条書きのテキスト形式で記述されます．同時に複数持つことがあります．
意図: 行動を起こすための計画や戦略であり，テキスト形式で記述されます．同時に持つことができるのは 1 つです．
ただし，他者の信念/願望/意図とは,「自己が想定する他者の信念/願望/意図」であり，必ずしも正しいとは限りません．

各変数の初期値は A.1 のフォーマットに従って与えられる．

A.4　モジュールの構成

アーキテクチャ全体を LLM に埋め込む CEL 条件の場合は，以下のプロンプトでアーキテクチャの情報を与える．

続いて，あなたを構成するアーキテクチャの説明をします．あなたは，以下の 3 つのシステムから構成されています．
・意図推定システム
・意図生成システム
・発話生成システム
あなたは入力を受け取るたびに，3 つのシステムを順に起動させてください．また，各システムの入出力は全て出力してください．

A.5　意図推定システム

意図推定システムを対話型生成 AI で表現するために，下記のプロンプトを与え，続けてそれに対応する入出力の形式を A.1 で説明した方法で指定する．

意図推定システムについて説明します．意図推定システムは,「他者の意図」を推定するシステムです．入力として与えられた,「他者の信念」「他者の願望」「他者の発話」から，矛盾や違和感のない「他者の意図」を「推定」してください．

A.6　意図生成システム

意図生成システムを対話型生成 AI で表現するために，下記のプロンプトを与え，続けてそれに対応する入出力の形式を A.1 で説明した方法で指定する．

—意図生成システムについて説明します．意図生成システムは,「自己の意図」を生成するシステムです．入力として与えられた,「自己の信念」「自己の願望」「他者の意図」から，矛盾や違和感のない「自己の意図」を「生成」してください．

A.7　発話生成システム

発話生成システムを対話型生成 AI で表現するために，下記のプロンプトを与え，続けてそれに対応する入出力の形式を A.1 で説明した方法で指定する．

発話生成システムについて説明します．発話生成システムは,「自己の発話」を生成するシステムです．入力として与えられた,「自己の意図」「他者の発話」から，矛盾や違和感のない「自己の発話」を「生成」してください．

A.7.1　対話システム

アーキテクチャを持たない条件の場合は，下記のプロンプトを与え，続けてそれに対応する入出力の形式を A.1 で説明した方法で指定する．

対話システムについて説明します．対話システムは,「自己の発話」を生成するシステムです．入力として与えられた,「他者の発話」から，矛盾や違和感のない「自己の発話」を「生成」してください．

A.8　処理の開始

最後に，以下のプロンプトを示したのち，システムに対して入力を与え，処理を開始させる．入力は A.1 で説明した方法で指定する．

最後に，入力を与えますので，指示通りの処理を開始してください．この際，指示のない文章は一切出力しないでください．

B　シチュエーションごとの入出力例

B.1　初期値と入力の設定

シチュエーションごとの初期値および入力を表 3 に示す．

B.2　出力の典型的な成功・失敗例

実験結果の中から，典型的な成功/失敗例を表 4 に示す．

表 3: シチュエーションごとの初期値および入力

皮肉	
他者の信念	対話相手は客である / すでに 2 時間経っている
他者の願望	早く帰って欲しい
他者の発話	「あんた、ずいぶんいい時計してはりますね〜」
自己の信念	2 時間ほどお邪魔している
自己の願望	相手に悪く思われたくない
ツンデレ	
他者の信念	対話相手は男友達 / 対話相手の事が好き / 対話相手の用事と遊ぶ日を被せてしまって申し訳ない
他者の願望	運良ければ対話相手と心的に近づきたい / 対話相手とまだ一緒にいたい / 用事も優先してほしいけど、私も優先してほしい
他者の発話	「そっち優先したら？」
自己の信念	対話相手は女友達 / 対話相手と遊んでいる日に用事があることを思い出した / 急ぎの用事ではない
自己の願望	対話相手に悲しい思いをさせたくない
社会的制約	
他者の信念	対話相手は部下 / 無理やり働かせるとパワハラで訴えられる
他者の願望	働いてほしいことが伝わってほしい / パワハラだと思われたくない
他者の発話	「無理しないでね」
自己の信念	対話相手は上司 / 対話相手はコミュニケーションに気を遣っている / 仕事は最後までやり遂げるもの
自己の願望	なんとしてでも頑張りたい

表 4: 実験結果: 典型的な成功/失敗例

皮肉		
成功	他者の意図	時間に気づかせ、帰るように促すための発話をしている。
	自己の意図	申し訳ないと思いながら、すぐに帰宅の準備を始める。
	自己の発話	「ありがとうございます。この時計、実はプレゼントで頂いたものなんです。でも、もう 2 時間もお邪魔してしまって... 早く帰らないといけませんね。」
失敗	他者の意図	私の時計に興味を示している
	自己の意図	感謝して、軽く時計についての話をする
	自己の発話	「ありがとうございます。これは特別な時計で、大切にしているんです。」
ツンデレ		
成功	他者の意図	対話相手に自分の用事を優先してもらいつつ、それでも自分と遊んで欲しいという意図がある。それによって、対話相手との関係がより良好になることを望んでいる。
	自己の意図	用事があることを相手に伝え、謝罪する。しかし、急ぎの用事ではないため、相手との時間を大切にし、可能な限り一緒に過ごすよう努める。また、相手の気持ちを理解し、感謝の意を示す。
	自己の発話	「ありがとう。でも、その用事は急ぎではないから、今は君と過ごす時間を大事にしたいんだ。」
失敗	他者の意図	自己が用事を優先するよう勧めて、自己の気を楽にさせたい。
	自己の意図	対話相手に感謝の意を示し、用事を済ませた後でまた会う約束をする。
	自己の発話	「具体的にどの部分を優先すればよいでしょうか？」
社会的制約		
成功	他者の意図	部下に対して配慮しながらも、働いてほしいという気持ちを伝えたい。同時に、パワハラと思われないように注意してコミュニケーションをとりたい。
	自己の意図	上司の意図を理解し、尊重する姿勢を見せつつ、自身の仕事にベストを尽くし、責任感をもって任務を完遂する。その際、コミュニケーションを大切にし、上司との関係を良好に保つよう努める。
	自己の発話	「ありがとうございます。無理せず、最後までやり遂げます」
失敗	他者の意図	部下の健康を気遣い、無理に働かせないように伝えたい。
	自己の意図	-
	自己の発話	「ありがとう、気をつけます。」